Problems in Replicating Multivariate Models in Quantitative Research. Health Psychology Update, 14(3), 10-16.

Problems in Replicating Multivariate Models in Quantitative Research

Ron Roberts
Kingston University
Department of Psychology
Penrhyn Road, Kingston, KT1 2EE
Email r.a.roberts@kingston.ac.uk

> *"it is a question of substituting signs of the real for the real"*
>
> *J. Baudrillard (Simulacra and Simulation 1981, p.2)*

In recent years psychological journals have devoted considerable space to discussing the relative merits of quantitative and qualitative research (e.g. Cooper and Stevenson, 1998; Morgan, 1998). Within these discussions, issues concerning the reliability and validity of findings have occupied a prominent place - in particular debates have examined the extent to which qualitative research methods can be said to legitimately address these aspects of the research process. The attention devoted to the legitimacy of qualitative methods from a scientific perspective however, has obscured the extent to which reliability and validity can also be problematic in quantitative research. In this article, the issue of replication from within the quantitative tradition will be examined, highlighting areas of potential difficulty and drawing attention to unresolved issues. The issue of the replicability of research findings goes to the heart of what we mean by science and any problems with it strikes at the heart of the legitimacy of psychology as a science and the knowledge claims it makes. In particular here the focus will be on the use of multivariate statistical models and the problems which arise in deciding whether these have been replicated. These models are widely used in health psychology, and involve the use of several variables in combination in order to account for a particular outcome. The specific illustrations and examples here are largely drawn from health psychology, though it is worth emphasising that the arguments presented are pertinent to practically all fields in psychology and to those disciplines which employ multivariate modelling to try and understand the world. We begin with a brief consideration of replication in more simpler contexts.

Problems in Replicating Multivariate Models in Quantitative Research. Health Psychology Update, 14(3), 10-16.

Statistical Significance, Effect Size and Replication

Whether a particular finding has been replicated is not simply a matter of deciding whether the same variable previously reported to produce a significant effect on a dependent variable (or been found to correlate significantly with another variable) has done so again. Regrettably this strategy has frequently been followed by health psychologists. Many prominent examples can be found - Social Cognition Models of health behaviours (Conner and Norman, 1995), the Type A Behaviour Pattern (Evans, 1998), and justification for psychological interventions in tackling health inequalities (e.g. Phillips and Pitts, 1998) have all received support on the basis of repeated statistically significant relationships being found between psychological variables and health outcomes. Recently the BPS's own Division of Health Psychology has announced an intention to collaborate with the Department of Health to undertake a systematic review of the role of self efficacy as part of a drive to tackle health inequalities - again on the basis of statistically significant relationships with health outcomes (For a detailed critique of the role of psychological interventions in addressing health inequalities see Carroll, Bennett and Davey Smith, 1993 and Macleod and Davey Smith, 2003).

Clarke-Carter's (2003) recent paper makes clear the shortcomings to this type of strategy in his consideration of effect size. If these are not considered unwarranted assumptions about the importance of an observed relationship may follow. First of all many effects go undetected because of an inadequate sample size - the power to detect them (i.e. a sufficiently large sample) wasn't there in the first place. Secondly an effect may be statistically significant but not practically, i.e. behaviourally or clinically significant.

In the present context reliance on statistical significance alone cannot provide a consistent benchmark as to whether an effect has been replicated because most studies employ different sample sizes. To be useful, the concept of replicability must address the size of any putative effect. As Clark-Carter (1997) has noted - the failure to address issues of statistical power - at least until recently has been widespread in psychology. Within my own branch of the discipline - health psychology, Conroy (1997) has similarly lamented the failure to perform power calculations when designing studies. Without consistent

measures of effect size, determining whether a finding has or has not been replicated is problematic to say the least. Where possible meta analyses have frequently been the tool of choice to address both the consistency and the magnitude of putative effects. But is this enough? If we are assessing the role of individual variables against specified outcomes, then appropriate meta analyses which examine the effect sizes in a selected series of published and unpublished studies are invaluable. Note however that discrepancies between the effect sizes reported in large randomised controlled trials and those estimated from meta-analyses have been found, suggesting the procedure is far from perfect (LeLorier et al. 1997) and as Field (2003) concluded current methods risk routinely inflating effect sizes.

Replicating multivariate models

1) Linear Regression Models

Greater problems however become apparent when more complex relationships between variables are considered. The enormous growth in computing power during the past two decades (Fox 1991) has enabled researches to examine more complex systems than had previously been possible. The recognition that health states, behaviours and beliefs for example are the product of complex interacting processes (Blane, Brunner and Wilkinson, 1996) has largely coincided with this growth in processing power. In the rush to exploit the increased processing power, researchers have frequently used multivariate techniques to explore and model the influences on health outcomes. Although reservations have occasionally been expressed about the use of these techniques in understanding health states and their psychosocial correlates (e.g. Haan, Kaplan and Syme 1989), some of the issues underlying their limitations have been overlooked. Thus, it is often the statistical models as a whole, notably linear regression models and structural equation models, rather than the role of individual variables that are under evaluation. Are there established criteria for establishing the replicability of these? I will argue below that the answer to this question is a resounding no.

Let us begin with the question of how to validate the results of a multivariate analysis. Suppose we set out to replicate a linear model where the outcome of interest (Y) is modelled in terms of the combined influence of three independent variables (x_1, x_2 and x_3).

In keeping with the desire to utilise the effect size of relationships, we will deal here with the standardised model in which the relationship between the independent and dependent variables is expressed in standard deviation units. Thus, the model is usually depicted as:

$Y= ß_1x_1+ß_2x_2+ß_3x_3 +ε$

Here beta (ß) represents the standardised measure of effect size for the relationship between each independent variable and the dependent variable Y. For the non-statistically minded amongst the readers, this can be considered as being akin to a correlation coefficient. The degree of error in the prediction is given by ε. What would replication of such a model involve?

In seeking to answer this, the first issue which arises concerns what data should be used to attempt the replication. This can be done in a cross-sectional study by splitting the initial sample into two and seeking to reproduce the pattern of results in both sub-samples. Alternatively, the researcher may seek to replicate using an entirely different sample or the same or different sample at an entirely different point in time. One could argue that if successful, this latter strategy would constitute a more robust replication. However, none of the strategies are without problems and quite different issues may arise, dependent upon which of them is followed.

Consider the first course of action - arguably the simplest - replication within the same sample. What would comprise successful replication here? A minimum requirement might be that the same variables attain statistical significance in each sub-sample. One could go even further and require that the effect sizes for each variable are comparable on both occasions - if so, this would entail the overall effect size for the model - the total explained variation (adjusted R-squared) remains the same. Certainly, if all these criteria are met then one could definitely say that the model has been replicated. What however if they are not - suppose only two of the three variables attain statistical significance, and only one of the three produces a comparable effect size in both samples with the overall R-squared different? In a situation such as this there is no common agreement - and the situation becomes more complicated still if one seeks replication of the model at a different point in time, as this introduces the possibility that the influence of certain variables has

changed in the interval since the initial study was undertaken. There are many examples of this - in the latter half of the 20th century for example, education has changed in its ability to predict occupational mobility (Goldthorpe, 1987), while the strength of the relationship between suicide and unemployment levels has also varied considerably in the post war period (Brenner 1979). So, failure to produce model equivalence may not indicate a lack of validity of the original model, and strict replication may just not be possible.

Consider the situation where replication is attempted with the same sample at a further point in time - results may indicate that the same predictors remain significant, but now the magnitude of each of the effects is reduced, with accordingly a reduction in overall effect size. One possibility that should not be overlooked is that insufficient variation may exist in the dependent variable during the later study compared to the earlier one. Whilst working on the Whitehall II study during the 1990's for example, Eric Brunner and myself found that a model of social mobility which predicted 50% of the variation (see Roberts, Brunner, White and Marmot, 1993), was only able to explain 4% when applied during a subsequent period.

Using calculations provided by Tabachnick and Fidell (1996) to determine the degree to which restricted variance in the dependent variable at a second time point can affect explained variance, we re-estimated the explained variance for the second period at 29% - thus finding that a majority (59.4%) of the reduction in explanatory power from the first period arose simply as a result of the reduced variability in the dependent variable during the second phase. Suitable transforms (see Tabachnick and Fidell, 1996) may be undertaken to assess the statistical significance of the difference in model strengths, where the variance in the dependent variable is comparable in both samples (a simple test for homogeneity of variance will suffice to determine this). Where it is not however, there are no hard and fast rules to follow which would enable one to decide what discrepancies in explained variance are tolerable in a successful replication.

If we consider logistic regression, the situation is complicated by a lack of agreed criteria for the predictive efficiency of a model, and uncertainty in some quarters about whether the primary aim of the analysis is to produce a statistically significant model, or

one where there is a strong association between the predictors and the outcome of interest (Menard,1995). Although logistic models are usually assessed for goodness of fit using chi-square techniques, and strength of association using the percentage of cases correctly classified by the model, analogous measures to the coefficient of determination (R^2) in linear regression are available with dichotomous dependent variables and permit estimation of the strength of a particular logistic model in the same manner as with conventional linear models (Menard, 1995).

2) Structural Equation Models

The issue of which criteria should inform acceptance of SEM model replication has seldom been considered in conventional printed literature. Miles and Shevlin's (2003) excellent introduction to the strengths of the technique goes no further than to indicate in a footnote to their article that replicability - a key scientific goal, is a problem. The issue has however received a good deal of attention in online discussion groups such as SEMNET (http://bama.ua.edu/archives/semnet.html). Structural Equation Modelling (SEM), sometimes referred to as path analysis or causal modelling brings its own set of problems regarding replication. A brief description of the basics of the method will help to clarify what some of these are, though I make no claim that the following discussion covers all the possible applications of SEM. SEM is a linear modelling technique for assessing goodness of fit (Dunn, Everitt and Pickles, 1993). Any particular structural model is described by a series of hypothesised linear equations and may also be represented diagramatically. In many applications the statistical question posed is whether the covariance matrix of the relationships specified in a given model departs significantly from another covariance matrix. The covariance matrix for comparison may belong to a competing theoretical model, or in the most basic case (the null model), comprise all possible relationships between the variables which are included in the model. Comparison between these is undertaken via a chi-square goodness of fit test. The degrees of freedom in comparing the specified model with the null model for example, comprise the number of relationships in the null model which do not appear in the specified model.

Some researchers using SEM have argued that the degree of explained variance in a

model is irrelevant in any potential replication - and that what is of paramount importance is the set of correctly specified structural relations between a set of variables (e.g. Hayduk, SEMNET 10 May 2002). This point of view has been challenged with the contention that if very little of the outcome is actually explained in a particular model, then the model itself is of little theoretical or practical importance (Barret, SEMNET 29 Oct 2002). Hayduk has in fact gone further and argued that replication itself should not be a major goal in structural equation modelling (Hayduk, SEMNET 8 May,14-15 May 2002). From these positions it is evident that controversy over replication is not confined to qualitative research.

Attempted replication of SEM models to date has employed cross-validation using multiple samples (see Cudeck and Brown 1983). However, this has largely been restricted to cross-sectional analyses, and therefore few SEM models to date have been replicated outside of the context in which they were first generated, and as with linear regression models, no agreement has been reached as to what actually constitutes replication of a structural model in the first place. The problems are similar, and concern indices of model fit in addition to those of the significance and effect size of individual pathways and the overall explained variance. A further major difficulty in interpreting the outcome of an attempted replication, is that even if analysis does decree that the model covariance matrices do not depart significantly from one another - it would be incorrect to infer from this goodness of fit test that we actually have a successful replication on our hands. Many analyses employing SEM have actually employed relatively small samples (sample size less than 200) where the power to reject the null hypothesis (and with it to detect failure to replicate) is quite small anyway. A vanishingly small number of studies using SEM have reported on the available power to reject the null hypothesis (*see* MacCallum, Browne and Sugawara 1996) no doubt in part because the power calculations are complex and are simply not available in many programs (Statistica is one exception).

Implications

The preceding arguments have I hope made it clear that replication of quantitative analyses is not the straightforward issue it might appear to be. No consensus exists regarding the criteria which must be employed for achieving successful replication of

multivariate models, and the question has received scant attention. Indeed, journals have appeared extremely reluctant to publish anything which draws reader's attention to the widespread failure to replicate the findings of complex statistical analyses. Continued avoidance can only play into the hands of those who question the validity of traditional scientific methods and what they purport to tell us about reality. If models which purport to represent reality are not subject to the requirement that they be replicated - then essentially, they are one-off ideas and as such cannot be considered to have survived any severe test of their validity. The distinction between science and non-science is not merely about the use of numerical methods per se to bolster an argument - it is about how an argument is constructed, how it is employed and how it may withstand potential criticism. A perusal of papers presented at some recent conferences might indeed lead one to wonder whether the aim of some research has become the building of structural models for their own sake, and whether teaching these techniques to new generations of students owes more to preparing them to survive in the academic rat race cum market place, suitably equipped to play the knowledge generation game, than actually to generate any findings of true worth. Several years ago, Dunn, Everitt and Pickles (1993), remarked that social and behavioural research literature was already full of work involving the fitting of increasingly complicated models, where they argued it was difficult to ascertain whether the authors had any realistic chance of rejecting their null hypothesis or discriminating between one model and another. There are few signs that the situation has improved since then.

Given the conventional view of scientific activity as expounded by Popper (1972), that a model or theory be considered valid to the extent that it enjoys repeated success in predicting phenomena, despite attempts to demonstrate its unworkability; the validity of regression and SEM models is inextricably bound up with their ability to yield repeatable findings. Less stringent philosophies of science than Popper's, which assert that scientific activity defines the limitations within which theories or models are applicable to the real world, would still place emphasis on repeatable testing (Dunbar, 1995). Thus, in employing multivariate statistical models, the issue of replication has considerable bearing on their

validity - and surely the primary aim of modelling must be to represent reality, i.e. to build models which incorporate sufficiently analogous features to those occurring in the real world (Barrow 1992; Cohen and Stewart, 1994). In the majority of instances in which statistical modelling has been employed in health psychology however, the resultant models have not been tested on new data, and have demonstrated very little validity (Roberts, Towell and Golding 2001). Thus, impressive as they sometimes appear to be, we must be clear about the information which they actually impart about reality. Statistical complexity does not translate into validity or 'truth'. What modelling strategy is followed, and the care with which inferences are drawn will in large part determine the value of the work to the wider scientific community. If the process of modelling in health psychology in particular and psychology in general is not to fall into disrepute and drag the discipline down with it, models must be constructed and interpreted with greater care than has thus far been evident.

Conclusion

We have highlighted here a number of points which must be addressed when considering the replication of multivariate statistical models. It is evident that no consensus exists regarding what is essential in any attempted replication, and indeed that the importance of replication itself is being challenged. Given this it might be helpful to consider a framework in which one can describe different degrees of replication. In SEM this could involve four levels - beginning with a good fitting model involving (A) the same variables. A higher level of replication would consist of (B) the same significant pathways between variables. A still greater level of replication (C) would see the same parameter estimates - with accordingly the same proportion of variance explained and finally (D) the same goodness of fit indices - dependent upon the power present in the study. This same framework could satisfactorily incorporate multiple regression models and confirmatory factor analysis. Failure to consider the issues discussed here entails serious problems for the credibility of scientific psychology and begs more questions about the aims of those who set the agenda within the discipline. Is it truth or power?

Problems in Replicating Multivariate Models in Quantitative Research. Health Psychology Update, 14(3), 10-16.

References

Barrow, J.D. (1992) Pi in the Sky: Counting, thinking and being. Oxford University Press.

Blane, D., Brunner, E. and Wilkinson, R. (Eds.) (1996) Health and Social Organisation. Routledge. London.

Carroll, D., Bennett, P. and Davey Smith, G. (1993) Socio-Economic Health Inequalities: Their Origins and Implications. Psychology and Health, 8, 295-316.

Clark-Carter, D. (1997) The account taken of statistical power in research published in the British Journal of Psychology. British Journal of Psychology, 88(1), 71-84.

Clark-Carter, D. (2003) Effect Size: the missing piece in the jigsaw. The Psychologist, 16(12), 636-638.

Cohen, J. and Stewart, I. (1994) The Collapse of Chaos. Penguin Harmondsworth

Conner, M. and Norman, P. (1995) (Eds.) Predicting Health Behaviour. Open University Press.

Cooper, N. and Stevenson, C. (1998) 'New Science' and Psychology. The Psychologist, 11(10) 484-485.

Cudeck, R. and Browne, M.W. (1983) Cross-validation of covariance structures. Multivariate Behavioural Research, 18, 115-126.

Dunbar, R. (1995) The Trouble with Science. Faber and Faber.

Dunn, G., Everitt, B. and Pickles, A. (1993) Modelling Covariances and Latent Variables using EQS. Chapman and Hall. London.

Evans, P. (1998) Coronary Heart Disease. In M. Pitts, and K. Phillips (Eds.) The Psychology of Health. (2nd Ed). Routledge. London.

Field, A.P. (2003) Can meta-analysis be trusted?The Psychologist, 16(12), 642-645.

Fox, J. (1991) Regression Diagnostics. Quantitative Applications in the Social Sciences. 79. Sage.

Goldthorpe, J.H. (1987) Social Mobility and Class Structure in Modern Britain (2nd Edition). Clarendon Press. Oxford.

Haan, M.N., Kaplan, G.A. and Syme, L. (1989) Some New Thoughts on Old Observations. In Bunker, J., Gomby, D.S. and Kehrer, B.H. (Eds.) Pathways to Health: The Role of Social Factors. Henry, J. Kaiser Family Foundation.

LeLorier, J., Gregoire, G., Benhaddad, A., Lapierre, J., and Derderian, F. (1997) Discrepancies between meta-analyses and subsequent large randomized, controlled trials. The New England Journal of Medicine, 337, 8, 536-542.

Lewis-Beck, M.S. (1980) Applied Regression: An introduction. Quantitative Applications in the Social Sciences. 22. Sage.

MacCallum, R. C., Browne, M. W., and Sugawara, H.M. (1996). Power analysis and determination of sample size for covariance structure modeling. Psychological Methods. 1(2), 130-149.

Macleod, J. and Davey Smith, G. (2003) Psychosocial factors and public health: a suitable case for treatment? Journal of Epidemiology and Community Health, 57(8), 565-570.

Menard, S. (1995) Applied Logistic Regression Analysis. Sage Quantitative Applications in the Social Sciences 106.

Morgan, M. (1998) Qualitative research...Science or pseudo-science? The Psychologist, 1(10), 481-483 & Postscript, 488.

Miles, J. and Shevlin, M. (2003) Navigating spaghetti junction. The Psychologist, 16(12) 639-641.

Phillips, K. and Pitts., M. (Eds.) (1998) The Psychology of Health (2nd Edition). Routledge. London.

Popper, K. (1972) Conjectures and Refutations. Routledge. London.

Roberts, R., Brunner, E., White, I. and Marmot, M. (1993) Gender Differences in Structure of Employment and Occupational Mobility in the British Civil Service. Social Science and Medicine, 37(12), 1415-1425.

Roberts, R., Towell, A. and Golding, J. (2001) Foundations of Health Psychology. Palgrave. Hants.

Tabachnick, B.G. and Fidell, L.S. (1996) Using Multivariate Statistics. Harper Collins.