A SAMPLING OF ANALYSES AND SOFTWARE USE FOR CLUSTER RANDOMIZED TRIALS OVER THE LAST DECADE

Elly J. H. Korendijk and Joop J. Hox
Utrecht University, Utrecht, The Netherlands
e.j.h.korendijk@uu.nl

In experimental research it is not always possible or desirable to randomize at the individual level; instead clusters of individuals are assigned to conditions. The clusters may be existing groups, like classes, families or communities; or may be established for the purpose of the trial, like therapy groups, where subjects within clusters are likely to respond more alike than subjects between clusters. Due to this dependency it is necessary to use hierarchical linear models, which are also referred to as multilevel models, when analyzing data of cluster randomized trials. A literature review over the last decade will summarize with which model data of cluster randomized trials have been analyzed and which software package has been used.

INTRODUCTION

In experimental research, it is not always feasible or desirable to assign subjects to conditions, and the intervention is delivered to subjects within identifiable groups or clusters. Assigning subjects within the same cluster to either the control or the experimental condition may be unfeasible due to costs, ethical reasons, or the risk of contamination (i.e., controls get information or learn from the subjects that are in the intervention condition, and hence are influenced by the intervention). In such situations it is better to randomize at the cluster level and hence these trials are called cluster randomized trials (Murray, 1998). For example in educational research schools or classes may be assigned to either the treatment or control condition, in organisational psychology it may be companies, in medical research it may be general practices and in health psychology it may be therapy groups.

In cluster randomized trials the subjects are referred to as the level one unit and clusters as the level two unit; level one units are said to be nested within the level two units and the data is said to have a hierarchical or multilevel structure. Since the clusters are identifiable, subjects within the same clusters share common features and are influenced by the same teacher, the same patron, or the same doctor or therapist. Furthermore, especially in therapy groups and classes, the subjects will mutually influence each other. In other words, in data of cluster randomized trials the assumption of independent observations is violated; subjects within a cluster will respond more alike compared to subjects in different clusters.

When designing cluster randomized trials or analyzing the resulting data, one should take into account the dependency of subjects within the same cluster. In the current paper we will focus on the analysis. Multilevel models, in some fields known as hierarchical models or mixed models, are designed for data with a nested structure, and hence take into account the dependency between subjects within the same cluster (e.g., Goldstein, 2003; Hox, 2010; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999).

The main risk of ignoring the hierarchical structure of data of cluster randomized trials, i.e. not analyzing the data with a multilevel model, is to find a significant intervention effect while in fact there is no such effect in the population. This is due to too small estimated standard errors of the intervention effect and hence too small p-values, plus too large number for the degrees of freedom (e.g. Murray, 1998, p. 6). However, when only a small number of clusters per condition are used, multilevel models are not performing adequately (e.g., Maas & Hox, 2005). In these situations a fixed effects model such as analysis of variance is more appropriate (e.g. Snijders & Bosker, 1999). [Note that from the perspective of power, it is not advisable to have only a few clusters per condition (e.g., Murray, 1998)].

Within the family of multivariate models the multilevel model is a relatively young member. In his textbook on cluster randomized trials Murray (1998) states "The last 25 years has witnessed dramatic improvements in the quality of the design and analysis of trials based on the allocation of identifiable groups to study conditions." (Murray, 1998, p. 13). The author gives credits to Donner and colleagues for a "steady stream of papers on the issues of analysis of facing

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php

group-randomized trials through the 1980s and 1990s." Since these early publications, a lot of research involving multilevel models in general and more specifically in the frame work of cluster randomized trials, has been published; a search in Google scholar within the last decade with the exact phrase 'cluster randomized trial' results in 1,220 hits and with 'group randomized trials' over 2,600.

In the mean time, software packages for the application of multilevel modeling became available. Some of these packages were designed as add-ons for general statistic packages, like SAS PROC MIXED and the GLIMMIX macro (Wolfinger, 1993). Other software packages are developed solely for multilevel modeling, like HLM (Raudenbush, Bryk, Cheong & Congdon, 2000) and MLwiN (Rasbash, Charlton, Browne, Healy & Cameron, 1997). In the general statistic software package SPSS multilevel modeling became available in release 11.0 under the term 'linear mixed models'. Finally, free online software packages like *R* have implemented multilevel modeling as well (Bliese, 2009).

Due to these developments, and the fact of implementing multilevel modeling in the curricula of social science education, the number of substantive researchers that recognize hierarchical data and are able to apply multilevel analysis is increasing.

Alongside the development of multilevel models, structural equation modeling (SEM) emerged, together with specific software as AMOS (SPSS Inc.) and LISREL (Scientific Software International). Complex data force statisticians to develop models and software that can handle the growing complexity. One of the results is software with the advantages of both the multilevel model and the structural equation model, e.g. Mplus (Muthén & Muthén, 1998).

Statistical models are developed when substantive research demands new models. However, from development to broad application is a long road. Donner, Brown and Brasher (1990) reported that less than 50% of cluster randomized trials published between 1979 and 1989 were analyzed adequately. Other studies, narrowed to a topic (e.g. school-based smoking interventions between 1975 and 1991 (Rooney & Murray, 1996)) or a specific field, e.g., primary prevention trials between 1990 and 1993 (Simpson, Klar & Donner, 1995), report similar percentages. The current study explores the adequacy of the analyses of data of cluster randomized trials and the software use in social sciences in the last decade (1999-2009).

METHOD

Within the different fields of social sciences (e.g., general psychology, behavioral psychology, developmental psychology, clinical psychology, cognition, sociology) the two electronically available journals with the highest impact factor are selected, plus two open source journals. By means of a thorough search within the volumes of 1999, 2004 and 2009 of the selected journals the papers reporting a cluster randomized trial are identified. When identified, some characteristics, as the applied analysis and software use, are noted.

To place the amount of cluster randomized trials within a broader perspective, within each year the percentage of papers reporting an experiment and the percentage of papers reporting a cluster randomized trial, are determined.

RESULTS

Applied analyses

In eleven of the 25 selected journals (see the appendix for all titles), not one cluster randomized trial is reported. In the journals that are reporting cluster randomized trials, the percentage of papers reporting an experiment is about equal in 1999 and 2009 (13.90% and 13.18% respectively) but is higher in 2004 (20.34%). Of these papers 6.12%, 14.12% and 10.53% are reporting a cluster randomized trial in 1999, 2004, and 2009 respectively. In absolute numbers, that is six papers reporting a cluster randomized trial in the volumes of 1999, 26 in 2004 and 21 in 2009.

In Table 1 the top four analysis techniques are considered to be appropriate for data of cluster randomized trials if applied in the correct situation, i.e. depending on the number of clusters in each condition. It can be seen that in 1999 only one study did apply an appropriate analysis, i.e. the one in the category 'Other analysis with correction for clustering'. In this study both conditions consisted of ten clusters and authors report to have applied multivariable analyses correcting for clustering. Though two studies report to have performed multilevel analysis, these studies had too few clusters per condition (i.e., three and nine respectively) for adequate application of this technique (e.g., Snijders & Bosker, 1999). The other three papers provided insufficient information (i.e., number of clusters was not given) to determine whether the applied analyses are appropriate.

Table 1. Number of different analyses applied on data of cluster randomized trials per year

Method	When Appropriate	1999	2004	2009
Multilevel Analysis	Many clusters	2 (na) [a]	1	4 2(na) [a]
General Equation Estimation	Many clusters		1	1
Analysis of Variance	Few clusters		9	5
Other analysis correcting for clustering	Depends	1	3	3
Aggregation	No limitations		3	
Ignored	Never		4	1
Unclear	Inadmissible reporting practice	3	5	5
Total		6	26	21

[a] na = with respect to the number of clusters not appropriate

In 2004 more than half of the papers apply the appropriate analyses. However, still 12 out of the 26 studies have applied an inappropriate analysis. Five years later, 13 out of 21 papers apply an appropriate analysis, but still eight papers report analyses that are inappropriate, because they show no evidence of taking the cluster structure into account in any way.

Three of the 2004 analyses are labelled by the authors of the current paper as ‘Other analysis with correction for clustering’. One of these papers reported to have used a technique developed by Donner & Klar (2000) for adjusting the variance. Another study accounted for clustering by using a formula for estimating the s.e. (standard error) developed by Altman (1991), and the third study used a technique suggested by Anderson and Ager (1978). The 2009 studies in this category report ‘adjusted for multilevel’, ‘adjusted for cluster randomization’ and ‘McNemar’s test’.

Software used

Many papers do not report the software used, although in 2009 this number is relatively small compared to the previous years; seven out of 21 in 2009 against three out of six in 1999 and 16 out of 26 in 2004. In Table 2 the reported software is listed per year. Various software packages are used, but general statistic software packages like STATA, SAS and SPSS are used most often. Software that can handle complex data (e.g. Mplus) is only used in the most recent year by three studies.

CONCLUSION

Although the proportion of cluster randomized trials of the published experimental research in 2009 is about equal to that of 1999, the proportion of researchers that apply appropriate techniques analyzing data of such trials has increased. Various techniques and software packages are used and most recently, there seems to be an increase of the use of software that has the advantage of both structural equation modeling and multilevel modeling.

In spite of the increasing proportion of appropriately applied analyses on data of cluster randomized trials, still 38% of the studies used inadequate analysis techniques or reported in such a way that the adequacy, or the lack of it, could not be determined. Both, the use of inadequate techniques and the lacunae in reporting about the technique and the software use, are quite troublesome; results are hard to interpret and impossible to replicate. In spite of the enormous amount of papers and books on the topic, many substantive social science researchers, reviewers and publishers, still seem unaware of the need for correcting for nesting when data come from cluster randomized trials.

Table 2. Software (frequency if reported by more than one paper) used per analysis by year

Method	1999	2004	2009
Multilevel Analysis		MLwiN	Mplus (2) not reported (2)
Multilevel Analysis (na[a])	SAS Proc Mixed ML3-E[b]		STATA & SPSS not reported
General Equation Estimation		STATA	STATA
Analysis of Variance		SPSS not reported (8)	SPSS (2) not reported (3)
Other analysis with correction for clustering		ACLUSTER SPSS not reported	STATA Mplus not reported
Aggregation		STATA (3)	
Ignored		not reported (4)	SPSS
Unclear	not reported (3)	SPSS not reported (3)	Statistica SAS & PROC TRAJ LISREL not reported (2)

[a] na = with respect to the number of cluster not appropriate
[b] ML3-E is an early version of MLwiN

Though not the focus of the current paper, some design issues draw our attention. For example, many studies have very small cluster level sample sizes; 60% of the studies reporting the cluster level sample size have less than 10 clusters. It is obvious that the power of such small sample sizes will rarely be sufficient. The gap between optimal design, power, required sample sizes, etcetera as postulated by various statisticians and methodologists (e.g., Hox, 2009; Maas & Hox, 2005; Moerbeek, Van Breukelen & Berger, 2000; Murray, 1998; Raudenbush & Bryk, 2002; Roberts & Roberts, 2005; Snijders and Bosker, 1999) and the practice as seen in the current sample of papers, show that substantive researchers are either unaware of design and power issues, or that in practice advised sample sizes are not feasible due to budgetary limits, lack of willingness of subjects to participate, or time.

To bridge the gap between theory and practice, statisticians and methodologists may need to publish more in journals for substantive research and not just in journals mainly read by fellow statisticians and methodologists. Furthermore, the today student is the tomorrow researcher; a student that receives interesting and good statistics and methodology classes may be a better future researcher. Teaching statistics is the challenge of statisticians.

REFERENCES

Arbuckle, J. L. (2006). Amos (Version 7.0) [Computer Program]. Chicago, IL: SPSS.

Altman, D. G. (1991). *Practical statistics for medical research,* London: Chapman & Hall.

Anderson, L. R., & Ager, J. W. (1978). Analysis of variance in small group research. P*ersonality and Social Psychology Bulletin*, *4*, 341-345.

Bliese, P. (2009). *Multilevel modeling in R* (2.3). Retrieved 15/2/2010 from http://cran.r-project.org/doc/contrib/Bliese Multilevel.pdf

Donner, A., Brown, K. S., & Brasher, P. (1990). A systematic review of non-therapeutic intervention trials employing cluster randomization, 1979-1989. *International Journal of Epidemiology*, *19*, 795-800.

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research.* London: Arnold.

Goldstein, H. (2003). *Multilevel Statistical Models.* London: Edward Arnold.

Hox, J. J. (2010). Multilevel analysis. Techniques and applications (2nd edition). NY: Routledge.

Jöreskog, K. G., & Sörbom, D. (2006). LISREL 8.8 for Windows [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.
Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel Modelling. M*ethodology, 1,* 86-92.
Murray, D. M. (1998). *Design and analysis of group-randomized trials*, Oxford: Oxford University Press.
Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.
Moerbeek, M., Van Breukelen, G. J. P., & Berger, M. P. F. (2000). Design issues for experiments in multilevel populations. *Journal of Educational and Behavioral Statistics*, *25,* 271-284.
R development core team (2007). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, http://www.R-project.org.
Rasbash, J., Charlton, C., Browne, W. J., Healy, M., & Cameron, B. (1997). *MLwiN Version 1*. Centre for Multilevel Modelling, University of Bristol.
Raudenbush, S. W., & Bryk, S. A. (2002). *Hierarchical linear models.* Thousand Oaks, CA: Sage.
Raudenbush, S. W., Bryk, S. A., Cheong, Y. F., Congdon, R. T., & Du Toit, M. (2004). *HLM 6*: *Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International, Inc.
Roberts, C., & Roberts, S. A. (2005). Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials*, *2*, 152-162.
Rooney, B. L., & Murray, D. M. (1996). A meta-analysis of smoking prevention programs after adjustment for errors in the unit of analysis, *Health Education & Behavior*, *23*, 48-64.
SAS (2008). Software 9.2. Cary, NC: SAS Institute.
Simpson, J. M., Klar, N., & Donner, A. (1995). Accounting for cluster randomization: A review of primary prevention trials, 1990 through 1993. *American Journal of Public Health*, *85*, 1378-1383.
Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modelling.* London: Sage.
SPSS for Windows, Rel. 11.0, (2001). Chigaco, IL: SPSS Inc.
StataCorp (2010). Stata Statistical Software. College Station, TX: StataCorp LP.
Wolfinger, R. D. (1993). *Fitting nonlinear mixed models with the new NLMIXED procedure*, Paper 287. Cary, NC: SAS Institute Inc.

APPENDIX

Table A. Journal titles with impact factor, 5-year impact, total number of papers, papers reporting experiments and papers reporting a cluster randomized trial (CRT) in 1999, 2004 and 2009

Journal Title	Impact factor 2008	5-year impact factor	Total number of papers	Total number of experi-ments	Total number of CRT
Age and ageing	3.052	3.118	252	33	9
AIDS and behavior	2.729	2.965	209	8	2
American J. of community psychology	1.198	2.313	139	4	2
American J. of sociology	2.808	5.046	104	1	0
Biological psychology	3.686	3.989	148	86	0
BMC Health service research[a]	1.168	1.947	204	5	5
BMC Public Health	2.029	2.256	430	18	1
British J. of psychology	1.671	2.276	99	42	0
British J. of sociology	1.475	2.173	81	0	0
Cognition	3.481	5.008	213	138	0
Drug and alcohol dependence	3.371	3.922	517	89	7
Health psychology	3.652	4.889	246	49	2
Human performance	1.159	2.165	53	6	0
Int. J. of behavioral development	0.983	1.721	165	14	0
J. of behavior therapy and experimental psychiatry	2.032	2.562	103	35	3

Journal Title	Impact factor 2008	5-year impact factor	Total number of papers	Total number of experi-ments	Total number of CRT
J. of clinical psychology	1.542	1.671	324	14	1
J. of clinical psychology in medical settings	0.427	–	105	13	5
J. of experimental social psychology	2.500	3.199	163	160	3
J. of memory and language	3.971	4.222	174	154	0
J. of occupational and organizational psychology	1.361	2.590	112	15	0
J. of pediatric psychology	2.895	3.446	195	21	3
Learning and instruction	1.435	2.446	105	23	6
Organizational behavioral and human decision processes	2.740	3.187	120	41	0
School psychology review	2.194	2.983	124	4	0

[a] This online open resource paper started in 2001, and hence this first volume was searched instead of a 1999 volume. – Indicates that there was no 5-year impact factor available.