Résumé

L'évaluation de l'efficacité des interventions : une exploration de deux méthodes statistiques

Mary T. Fox, Angela Cooper Brathwaite et Souraya Sidani

La formule de mesures répétées est souvent utilisée pour évaluer l'efficacité des interventions. Selon cette formule, les résultats sont mesurés à plusieurs reprises, soit avant et après la mise en œuvre du plan d'intervention. Les données peuvent être analysées à l'aide de deux méthodes statistiques : l'analyse de la variance fondée sur le mesurage répété (RM-ANOVA) et le modèle linéaire hiérarchique (HLM). Les auteures offrent un aperçu des modèles statistiques propres à la RM-ANOVA et au HLM et discutent des forces et des limites de chacun d'eux tout en suggérant que ces deux méthodes sont complémentaires dans une démarche visant à mesurer l'efficacité des interventions.

Mots clés : analyse de la variance

CJNR 2004, Vol. 36 N° 3, 20–30

Evaluating the Effectiveness of Interventions: Exploration of Two Statistical Methods

Mary T. Fox, Angela Cooper Brathwaite, and Souraya Sidani

Repeated measures designs are often used to evaluate the effectiveness of interventions. In these designs, the outcomes are measured on several occasions before and after implementation of the intervention. Two statistical methods, the repeated measures analysis of variance (RM-ANOVA) and hierarchical linear models (HLM), can be used to analyze the data. The authors provide an overview of the statistical models underlying RM-ANOVA and HLM and discuss the strengths and limitations of each. They propose that the 2 methods are complementary in determining the effectiveness of interventions.

Keywords: research design, treatment outcome, outcome assessment (health care), analysis of variance, regression analysis, linear models, longitudinal studies, nursing research

Background

The provision of empirical evidence regarding the effectiveness of interventions is critical for decision-making in clinical, administrative, and educational nursing practice. Appropriate implementation requires knowledge about the extent to which an intervention is beneficial. Such knowledge is usually derived from the results of studies that evaluated the effects of an intervention on the intended outcomes. Intervention evaluation studies tend to use repeated measures designs. A repeated measures design is one in which the same outcome variables are measured for each participant at several points in time (Daniel, 1999; Keselman, Algina, & Kowalchuk, 2001). The study involves one or more groups. The points in time are selected to represent the participants' status before and after implementation of the intervention. The statistical analysis is aimed at determining the extent to which the participants' standing on the outcomes changed as a result of the intervention.

Two statistical methods, repeated measures analysis of variance (RM-ANOVA) and hierarchical linear models (HLM), can be used to analyze the data obtained from intervention studies with repeated measures designs. These methods provide different yet complementary information about the effectiveness of the intervention. The findings of RM-ANOVA

indicate whether the mean scores on the outcome differed across occasions of measurement, within the same group or between groups (e.g., experimental and control groups). These findings, presented at the group level, are informative. However, they may not be sufficient to guide decision-making in practice; nurses need to know the percentage of participants who demonstrated the anticipated pattern of change in the intended outcome (Jacobson & Traux, 1991; LeFort, 1993; Sidani & Braden, 1998). HLM is a complementary data analytic method. It generates findings showing the pattern of change in the outcome for individual participants and the percentage of participants who demonstrated the anticipated improvement.

In this paper, we briefly review the statistical models underlying RM-ANOVA and HLM and discuss the strengths and limitations of each method in intervention evaluation research. We then present an empirical example to illustrate the complementary nature of the results obtained using the two methods; the results provide a comprehensive depiction of the effectiveness of the intervention using both methods.

Repeated Measures Analysis of Variance (RM-ANOVA)

RM-ANOVA is used to assess the effectiveness of an intervention in studies with repeated measures of the outcomes. The outcomes are measured at pre-test, post-test, and follow-up intervals in either the same group of participants who received the intervention, or two or more groups of participants who did or did not receive the intervention — that is, experimental or control groups, respectively. In experimental research, it is assumed that the intervention effects show up as a change of a constant value in the participants' post-test outcome scores (Lipsey, 1990; Raudenbush & Bryk, 2002). Therefore, the data analysis focuses on determining whether the mean scores on the outcome differ between groups and/or across occasions of measurement.

RM-ANOVA is a "multivariate inferential statistical procedure that is used to compare performance on a single dependent variable at three or more points in time" (Norwood, 2000, p. 364). The statistical model underlying RM-ANOVA is an extension of the general linear model. It tests for group, time, and group-by-time interaction effects, depending on the number of groups included in the study. The appropriate F ratio is computed to determine whether each effect is statistically significant. A statistically significant group effect indicates that the mean scores on the outcome differ between the experimental and control groups; however, it gives no clear indication of the measurement occasion(s) at which the between-group difference occurred. Post hoc analyses using independent sample t tests are conducted to compare the means of the two groups at

each point in time. A statistically significantly time effect implies that the mean scores on the outcome for the total sample differed across occasions of measurement, but there is no clear indication of when exactly the difference was noticed. Post hoc analyses using paired *t* tests are conducted to compare the means observed at the different occasions of measurement. In the latter case, the mean comparisons between occasions are done within each group as well as for the total sample (Green, Salkind, & Akey, 1999). A statistically significant group-by-time interaction effect indicates that the mean scores on the outcome differed between groups and across occasions of measurement. Post hoc comparisons are conducted to identify the point(s) in time at which the groups differed (Green et al.; Littell, Henry, & Ammerman, 1998).

The *F* ratio used in the RM-ANOVA and the *t* test used in the post hoc comparisons examine differences in the outcome at the group level. The results are presented for the "average" participant (Barlow, 1996) in terms of an "average" amount of change in the outcome between adjacent points in time (Hartman, Stage, & Webster-Stratton, 2003; Wu, Clopper, & Wooldridge, 1999). The emphasis on the group-level analysis and the mean change in the outcome may not be sufficiently informative to guide practice (Jacobson & Truax, 1991) and may be potentially misleading (Francis, Fletcher, Stuebing, Davidson, & Thompson, 1991). The *F* ratio and the post hoc *t* tests, and their associated results, do not provide any information either on the variability in the outcome achieved by the participants or on the pattern of change in the outcome across occasions of measurement. The participants may have differed in their responses to the intervention; some may have fully benefited, others not benefited at all, and others benefited to varying degrees (Brown, 2002; Sidani & Epstein, 2003).

In the statistical model underlying the RM-ANOVA *F* ratio and the post hoc *t* tests, the individual variability in the outcome is reflected in the within-group variance, which is considered the error term (Francis et al., 1991). Increased within-group variance reduces the statistical power to detect significant differences in the outcome between groups and/or across occasions of measurement, which in turn increases the likelihood of committing a type II error — that is, erroneously concluding that the intervention was not effective (Lipsey, 1990). Therefore, it is critical to know how many participants demonstrated the anticipated improvement in order to determine the benefits derived from the intervention (Jacobson & Truax, 1991; LeFort, 1993).

The results of RM-ANOVA and post hoc comparisons indicate that the mean scores differed over time, suggesting a certain amount of change in the group's level on the outcome. These findings do not provide information on the pattern of change — that is, the direction

and magnitude of the change in the outcome (Bryk & Raudenbush, 1987). Trend analysis can be performed to determine the pattern of change in the outcome; however, this analysis focuses on the average change observed for the group, not individual variability in the pattern of change (Francis et al., 1991; Wu et al., 1999). Evidence regarding the number of participants who changed and their patterns of change not only is clinically relevant but also may minimize the potential for drawing inaccurate conclusions regarding the effectiveness of the intervention.

Hierarchical Linear Models (HLM)

HLM, also known as growth curve or individual regression analysis, is a statistical technique that can be used to analyze individual differences in the pattern of change in the outcome (Floreck & De Champlain, 2001; Willett, Singer, & Martin, 1998). Its results complement those obtained from RM-ANOVA; the results derived from HLM describe the direction and magnitude of change in the outcome for each participant and indicate whether the pattern of change differed between the study groups (Raudenbush & Bryk, 2002; Warschausky, Kay, & Kewman, 2001). The assumption underlying HLM is that change is a continuous process, which is best described by a trajectory rather than a series of discrete alterations observed at fixed points in time (Francis et al., 1991; Wu et al., 1999). The focus is on modelling, describing, and explaining the variability in the trajectory or pattern of change.

To clarify the statistical model of HLM, we present it in two steps, the first describing the analysis conducted at the individual level, the second explaining the analysis conducted at the group level. The second step is performed if two or more groups are included in the study.

In the first step, the analysis estimates the pattern of change in the outcome measured before and after implementation of the intervention for each participant (Lipsey & Cordray, 2000). In this analysis, the data should be entered for each participant by occasion of measurement, as detailed by Raudenbush and Bryk (2002) and Sidani and Lynn (1993). A time variable is created to reflect the different points of measurement, which are assigned consecutive numeric values. The outcome variable is created to reflect the score obtained at each point of measurement. The outcome variable is regressed on the time variable, for each individual. The regression equation is expressed as $Y_i = B_{oi} + B_{1i}$ (Time) + error. Y_i represents the outcome variable for each participant. B_{oi} is the intercept; it describes each individual participant's value on the outcome at pre-test. B_{1i} is the slope; it describes the pattern — that is, the direction and magnitude — of change in the outcome for each participant. Time

represents the variable reflecting the different points of measurement. Error is the random error of prediction.

Each participant's slope is examined for its direction. A negative slope indicates that the participant's level on the outcome decreased, whereas a positive slope implies that it increased over time. Also, the standardized slope is evaluated for its magnitude. A slope of zero indicates that no change occurred in the outcome over time. A slope of less than .30 indicates that the participant's level on the outcome showed a small, gradual change over time. A slope between .30 and .60 indicates that the participant's level on the outcome showed a moderate but steady change. A slope greater than .60 indicates a large, rapid change in the outcome. Visual examination of the individual regression lines is highly recommended, to determine whether some participants exhibited non-linear, such as inverted ∩ or S-shaped, patterns of change. For these participants, the regression model should incorporate the appropriate terms (e.g., polynomials) that account for non-linearity (Kleinbaum, Kupper, & Muller, 1988; Warschausky et al., 2001).

In the second step of HLM, the analysis is intended to determine the extent to which the pattern of change in the outcome varied between the groups of participants. This is done with regression-type analyses. The individual participants' slopes (B_{1i}) that were obtained in the first step are regressed onto the variable(s) hypothesized to influence the pattern of change in the outcome. These may include: (1) the participant's level on the outcome measured at baseline or pre-test, which is represented by the intercept (B_o) obtained in the first step — the baseline outcome level is usually included in the regression analysis when significant inter-individual differences are observed at baseline; (2) the group to which the participant was assigned or the dose of the intervention to which the participant was exposed, as suggested by Sidani (1998). The latter variable is used when a dose-response analysis is conducted. The regression weights or beta parameters associated with each variable are examined for magnitude, direction, and statistical significance, as they indicate the extent to which these variables influenced the pattern of change in the outcome (Raudenbush & Bryk, 2002; Tate & HoBanson, 1993; Warschausky et al., 2001).

The results of HLM point to the direction and magnitude of change in the outcome exhibited by the participants from pre-test to post-test or follow-up. Variability in the pattern of change is acknowledged and considered of importance and interest. The number and percentage of participants who showed no change or different patterns of change are reported and may indicate the extent to which the intervention was beneficial. Intervention benefits can be inferred if a large percentage of par-

ticipants who received it showed the anticipated direction and magnitude of change in the outcome.

Illustrative Example

The data used in the following illustrative example were obtained from a study that evaluated the effectiveness of a staff-development educational program aimed at enhancing nurses' cultural knowledge and competence (Cooper Brathwaite, 2004). The program consisted of five sessions during which the theory and principles underlying cultural competence were discussed and the skills required for providing culturally competent care were reinforced. The program was offered to 76 public health nurses.

A one-group repeated measures design was used to evaluate the effects of the program. This design, which ascertains pattern of change over time in the same group of individuals and change within each individual under control and post-intervention conditions (Burns & Grove, 2001), was selected for several reasons. By not having separate treatment and control groups, it avoided dissemination of the intervention beyond the treatment group and compensatory rivalry among nurses working at the same site (Cooper Brathwaite, 2004). Also, the design minimized the potential for selection bias since the intervention was not given to nurses working at one site and withheld from those working at another (Cooper Brathwaite).

The outcome data were collected at four occasions separated by equal time intervals of 3 months. The first two occasions (time 1: first pretest; time 2: second pretest) represented the control condition. Time 3 data were obtained following implementation of the program. Time 4 data reflected the 3-month follow-up, which was used to determine the sustainability of the program's effects. The outcome variables were measured with self-report instruments that demonstrated acceptable reliability and validity. Cultural knowledge was measured using the Cultural Knowledge Scale (CKS) adapted from the cultural knowledge and cultural efficacy subscales developed by Campinha-Bacote (1999) and Bernal and Froman (1993), respectively. Cultural competence was measured using Campinha-Bacote's Inventory for Assessing the Process of Cultural Competence (IAPCC).

The RM-ANOVA was used to determine whether the group's mean scores on each outcome differed across occasions of measurement. A significant time effect [$F(3, 69) = 142.02$, $p < .01$)] was found for CKS. The post hoc comparisons showed that: (a) the mean score at time 1 ($3.76 \pm .26$) did not differ significantly from the mean score at time 2 ($3.77 \pm .27$); (b) the mean score at time 2 was significantly different from the mean score at time 3 ($4.57 \pm .34$); and, (c) the mean score at time 3

did not differ from that observed at time 4 (4.58 ± .39). Similarly, a significant time effect [$F(3, 69) = 118.87, p < .01$] was found for IAPCC. The results of the post hoc comparisons indicated that the mean scores differed between time 1 (2.87 ± .23) and time 2 (2.82 ± .18), time 2 and time 3 (3.38 ± .34), and time 3 and time 4 (3.51 ± .37).

The data were also analyzed using HLM in order to describe the pattern of change in each outcome within the group. The slope was estimated for each participant. For CKS, the estimated slope values ranged from .02 to .59, with a mean of .32 (± .14). Specifically, four participants (5.6%) had a slope close to zero (0 to .10), reflecting no gain in their cultural knowledge over time. For 47.2% of the participants, the slope varied between .11 and .30, reflecting a small, gradual gain in knowledge. For the remaining 47.2% of the participants, the slope ranged from .31 to .60, implying a moderate but steady gain in cultural knowledge. For IAPCC, the estimated slope value ranged between .00 and .53, with a mean of .23 (± .12). About 18.1% of the participants had a slope value of .00 to .10; 55.5% had a slope value of .11 to .30; and 26.4% had a slope value of .31 to .60.

The findings of the RM-ANOVA indicated that (a) the intervention was effective in increasing the nurses' level of cultural knowledge and competence (evidenced by the significant difference between the mean scores at times 2 and 3), and (b) the level of knowledge attained at posttest was maintained at 3-month follow-up, while the level of competence continued to increase at follow-up, as hypothesized. Although these results support the effectiveness of the intervention, they do not inform us of how many participants showed the anticipated improvement in the outcomes. The findings of the HLM clarified that approximately half of the participants increased their cultural knowledge and about a quarter demonstrated improvement in their level of cultural competence.

Conclusion

The clinical usefulness and applicability of results derived from intervention studies rests on the researcher's ability to provide evidence on the nature of change that results from an intervention. The analysis of longitudinal data derived from intervention studies has traditionally been conducted using RM-ANOVA. As such, the description of change has been chiefly limited to whether or not there was a change for the average participant from the target population. Although this is valuable information, it is insufficient as a basis for clinical, educational, and administrative decisions in nursing. Knowing the extent to which participants changed and their specific patterns of change is requisite for evidence-based decision-making. This level of information is not provided by RM-ANOVA.

Therefore, nurse researchers need to expand their repertoire of data-analytic approaches. HLM is an approach that researchers can use to complement RM-ANOVA in the analysis of longitudinal data derived from intervention studies.

Used together, RM-ANOVA and HLM equip researchers to provide more complete information on the effectiveness of interventions. RM-ANOVA can provide information on differences in change between groups. HLM can complement this level of information by describing the pattern of change, how various subgroups and/or individuals of the population changed, the proportion of individuals who changed, and the ways in which they changed in response to the intervention. For example, the information presented in the illustrative example using RM-ANOVA indicated that the educational program was effective, on average, in increasing public health nurses' cultural knowledge and competency. Although this information can be helpful to the nurse administrator in deciding whether to implement the educational program, it cannot be used to infer the efficacy or utility of the program. However, using the additional information obtained from the HLM analysis, the administrator can anticipate that the educational intervention will benefit approximately half of the public health nurses in terms of increasing their cultural knowledge and approximately one quarter in terms of increasing their cultural competence. Moreover, the nurse administrator can use the results from the HLM analysis to anticipate that 47% of the nurses who attend the educational program will have a moderate steady increase in cultural knowledge and 26% will have a moderate steady increase in cultural competence over a 3-month period. Furthermore, this additional information can be extended to estimate the cost effectiveness of the program.

References

Barlow, D. H. (1996). Health care policy, psychotherapy research, and the future of psychotherapy. *American Psychologist, 51,* 1050–1058.

Bernal, H., & Froman, R. (1993). Influences on the cultural self-efficacy of community health nurses. *Journal of Transcultrual Nursing, 4*(2), 24–31.

Brown, A. J. (2002). Nursing intervention studies: A descriptive analysis of issues important to clinicians. *Research in Nursing and Health, 25,* 317–327.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101*(1), 147–158.

Burns, N., & Grove, S. K. (2001). *The practice of nursing research*. Toronto: W. B. Saunders.

Campinha-Bacote, J. (1999). A model and instrument for addressing cultural competence in health care. *Journal of Nursing Education, 38*(5), 203–207.

Cooper Brathwaite, A. (2004). *Evaluation of a cultural competence educational intervention*. Unpublished doctoral dissertation, University of Toronto.

Daniel, W. (1999). *Biostatistics: A foundation for analysis in the health sciences* (7th ed.). Toronto: John Wiley.

Floreck, L. M., & De Champlain, A. E. (2001). Standardized patients – will the questions never end? *Academic Medicine, 76*(10), S93–S95.

Francis, D. J., Fletcher, J. M., Shuebing, K. K., Davidson, K. C., & Thompson, N. M. (1991). Analysis of change: Modeling individual growth. *Journal of Consulting and Clinical Psychology, 59*(1), 27–37.

Green, S., Salkind, N., & Akey, T. (1999). *Using SPSS for Windows: Analyzing and understanding data* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Hartman, R. R., Stage, S. A., & Webster-Stratton, C. (2003). A growth curve analysis of parent training outcomes: Examining the influence of child risk factors (inattention, impulsivity, and hyperactivity problems), parental and family risk factors. *Journal of Child Psychology and Psychiatry, 44*(3), 388–398.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*(1), 12–19.

Keselman, H. J., Algina, J., & Kowalchuk, R. K. (2001). The analysis of repeated measure designs: A review. *British Journal of Mathematical and Statistical Psychology, 54,* 1–20.

Kleinbaum, D. G., Kupper, L. L., & Muller, K. E. (1988). *Applied regression analysis and other multivariable models* (2nd ed.). Boston: PWS-KENT.

LeFort, S. M. (1993). The statistical versus clinical significance debate. *Image: Journal of Nursing Scholarship, 25*(1), 57–62.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Thousand Oaks, CA: Sage.

Lipsey, M. W., & Cordray, D. S. (2000). Evaluation methods for social intervention. *Annual Review of Psychology, 51*, 345–375.

Littell, R. C., Henry, P. R., & Ammerman, C. B. (1998). Statistical analysis of repeated measures data using SAS procedures. *Journal of Animal Science, 76*(4), 1216–1231.

Norwood, S. L. (2000). *Research strategies for advanced practice nurses.* Upper Saddle River, NJ: Prentice Hall.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Sidani, S. (1998). Measuring the intervention in effectiveness research. *Western Journal of Nursing Research, 20*(5), 621–635.

Sidani, S., & Braden, C. J. (1998). *Evaluating nursing interventions: A theory driven approach*. London: Sage.

Sidani, S., & Epstein, D. R. (2003). Enhancing the evaluation of nursing care effectiveness. *Canadian Journal of Nursing Research, 35*(3), 26–38.

Sidani, S., & Lynn, M. R. (1993). Examining amount and pattern of change: Comparing repeated measures ANOVA and individual regression analysis. *Nursing Research, 42*(50), 283–286.

Tate, T. L., & HoBanson, J. E. (1993). Analyzing individual status and change with hierarchical linear models: Illustration with depression on college students. *Journal of Personality, 61*(2), 181–206.

Warschausky, S., Kay, J. B., & Kewman, D. G. (2001). Hierarchical linear modeling of FIM instrument growth curve characteristics after spinal cord injury. *Archives of Physical Medicine Rehabilitation, 82,* 329–334.

Willett, J. B., Singer, J. D., & Martin, N. C. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations. *Development and Psychopathology, 10,* 395–406.

Wu, Y. W. B., Clopper, R. R., & Wooldridge, P. J. (1999). A comparison of traditional approaches to hierarchical linear modeling when analyzing longitudinal data. *Research in Nursing and Health, 22,* 421–432.

Authors' Note

Comments or inquiries may be directed to Mary Fox, Director, Collaborative Research Program, Rehabilitation and Long-Term Care, Nursing Services Administration, Baycrest Centre for Geriatric Care, 3560 Bathurst Street, Toronto, Ontario M6A 2E1 Canada. Telephone: 416-785-2500, Ext. 2714. Fax: 416-785-2501. E-mail: mfox@baycrest.org

---

Mary T. Fox, RN, MSc, is a PhD candidate in the Faculty of Nursing, University of Toronto, Ontario, Canada; Director, Collaborative Research Program, Rehabilitation and Long-Term Care, Baycrest Centre for Geriatric Care, Toronto; and a fellow of the Canadian Institutes of Health Research. Angela Cooper Brathwaite, RN, MN, CHE, PhD, is Manager, Public Health Nursing and Nutrition, Durham Region Health Department, Whitby, Ontario; and Lecturer, Faculty of Nursing, University of Toronto. Souraya Sidani, RN, PhD, is Associate Professor, Faculty of Nursing, University of Toronto.