The Aggregation of Single-Case Results using Hierarchical Linear Models

Wim Van den Noortgate and Patrick Onghena

Katholieke Universiteit Leuven, Belgium

.

Abstract

To investigate the generalizability of the results of single-case experimental studies, evaluating the effect of one or more treatments, in applied research various simultaneous and sequential replication strategies are used. We discuss one approach for aggregating the results for single-cases: the use of hierarchical linear models. This approach has the potential to allow making improved inferences about the effects for the individual cases, but also to estimate and test the overall effect, and explore the generality of this effect across cases and under different conditions.

Keywords: single-case; hierarchical linear model; replication; aggregation

Introduction

Single-case experimental designs are used to evaluate the effect of one or more treatments on a single case. The case may be a subject or another single entity that forms the research unit, such as a school or a family. This entity is repeatedly observed, over the levels of one or several manipulated independent variables (Onghena, 2005). In the most basic design, the AB-phase design or interrupted time series design, the case is observed repeatedly during a first phase (A), typically a baseline phase before an intervention takes place, and in a second phase (B) after or during an intervention. To evaluate the effect of the intervention, scores in both phases are compared.

Single-case designs have a long history in behavioral science (Ittenbach & Lawhead, 1997), but the last decades, single-case methodology has further been elaborated, aiming at improving the internal validity of the conclusions. For instance, reversal phase designs (e.g., an ABAB-design) or alternation designs with rapidly alternating conditions (e.g., an AABBBABAABB-design) rather than a simple AB-phase design may be used in order to assess or control statistically for the effect of history, maturation or other time-related confounding variables. The effect of such confounding variables may further be controlled by means of randomization while setting up the study, for instance by randomly assigning measurement occasions over treatments or randomizing the time of intervention (Edgington, 1996).

Although group designs receive much more attention in methodological courses and handbooks, in the last decades there has been renewed interest in single-case designs, especially in behavior modification and clinical psychology (Barlow & Hersen, 1984. Kazdin, 1982), neuropsychology (Caramazza, 1990), psychopharmacology (Cook, 1996), and educational research (Kratochwill & Levin, 1992). The popularity of the designs is also reflected in the relatively large number of articles published in the *Behavior Analyst Today*

that discuss or apply a variety of single-case designs (about twenty between 2001 and 2006). Single-case designs indeed are very attractive in several situations (Franklin, Allison, Gorman, 1997; Onghena, 2005). Single-case studies may be relatively easy to set up and are much less expensive than large-scale group-comparison studies. This makes the designs also attractive for practitioners, who want to get a first insight into the effect of a treatment.

An additional strength of single-case designs is that, in contrast to group designs that give insight into the average effect of a treatment, they give an in-depth insight into the behavior of one single case. Especially in clinical settings, the research indeed often focuses on the effect of a treatment for a specific case.

Finally, since only a single case is investigated, the design often allows making a large number of repeated observations, enabling a detailed study of the evolution of the behavior.

Single-case designs thus are (initially) aimed at drawing valid conclusions regarding one entity. Sometimes, for instance in applied clinical settings, the primary interest may indeed be in this single entity, since it concerns a case that presented itself with a problem to solve. Moreover, the problem presented by the client may be relatively rare. Yet, the question often arises what can be learned from this case for other cases, or if and to what degree the results of the case can be generalized to other similar cases. An important drawback of single-case experimental studies is that the results are in principle restricted to the cases that were studied. In situations for which it seems reasonable that the effect is common for a whole population, the results of a single-case study may be informative for other entities of this population, but such a generalization cannot be made based on statistical grounds.

A natural way to explore the generalizability of single-case results is replication over other entities. According to Barlow and Hersen (1984), this can be done by a) direct replication, which is a replication of the study by the same investigator but on other entities, b) systematic

replication, replicating the study varying one or more factors such as the characteristics of the setting, the experimenter or the treatment, and c) clinical replication, which is a more advanced replication by the same investigator of a treatment package containing two or more distinct treatment procedures on a series of clients presenting similar problems. Such replications can be carried out sequentially, for instance every time a client enters a clinical setting. Alternatively, simultaneous replication can be part of the study design. In a multiple baseline across participants design for instance, several cases are investigated simultaneously by means of an AB-phase design (Ferron & Scott, 2005).

While group designs typically give information about the effect of a treatment on 'an average case', and single-case designs give information about the effect on a specific case, a set of single-case studies combines the strengths of both designs. It offers the opportunity to draw conclusions about specific cases, but also about an average or typical case, as well as exploring systematically the conditions under which an effect will or will not occur. In this article, we want to describe a way of aggregating the results from single cases in a quantitative and systematic way, maximally exploiting the information that is available in the data: the use of hierarchical linear models. The approach will be introduced by means of an elaborated example.

An example

In a recent number of the *Journal of Early and Intensive Behavior Intervention*, Lawson and Greer (2006) use a multiple baseline design with middle school students with academic delays, to study the effects of writer immersion and viewing the effect of their writing on responses from readers on the quality of writing. Participants were seven 9th graders, diagnosed with behavioral and learning disabilities. In a first experiment, three students were asked in several sessions to write an essay describing a simple picture, such that a reader

could reconstruct the picture without seeing it. Initially, in the baseline phase, they were not given any feedback. In a second phase, the teacher gave them written and verbal praise for correct structural components (grammar and spelling), and corrections for incorrect responses, as well as verbal feedback on the content of the descriptions. In a final phase, the writer immersion phase, all communication was done in written form. Feedback in this phase included comments on the structural components, as well as viewing the effect their writing had on the reader who tried to draw the picture based on the description, being blind for the objectives or the setup of the study.

In both intervention phases, the pupils were asked to revise their essay after receiving the feedback. Cycles of feedback and revision continued until the descriptions were written correctly and resulted in correct drawings when used as instructions for a reader. In Figure 1, the results are shown for one of the dependent variables, more specifically, the structural components of the text, evaluated in each session prior to feedback. The variable equals the number of correct responses to spelling, punctuation, capitalization, word choice, and sentence structure, divided by the total number of opportunities to respond within each essay, multiplied by 100 %.

Insert Figure 1 about here

The second experiment was a systematic replication of the first experiment. To assess the effect of writer immersion, without being preceded by the traditional approach of simple teacher editing, the effect of writer immersion was now investigated for four new cases, but without the second phase. Results are displayed in Figure 2.

Insert Figure 2 about here

It is clear from Figure 1 that for the cases of the first experiment, scores in the teacher feedback phase and especially in the writer immersion phase are substantially higher than the scores in the baseline phase. The large effect of writer immersion that was found in the first experiment however could be a cumulative effect of teacher editing alone and writer immersion. The second experiment gives evidence for a substantial effect of writer immersion, even when not preceded by the simple teacher editing phase.

This is also found if for each student the mean percentages for the phases are calculated and compared, as is done by the authors. For the first subject (Student A from Figure 1) for instance, the mean percentage was 30 % for the first phase, as compared to 73.25 for the second and 83.33 % for the third phase, a rather impressive difference of 40.25 % and 53.33 %! Unfortunately, large effect magnitudes are relatively rare in behavioral science (Cohen, 1988; Pillemer, 1984). Therefore drawing conclusions based on visual inspection of a graphical display or of summary statistics may be risky since small effects are difficult to see with the naked eye and therefore may remain undetected (Edgington & Onghena, 2007), although there is also evidence that based on visual inspection, it is too easily concluded that there is an effect (Matyas & Greenwood, 1990).

Considering this, it is recommended to supplement the visual analysis of the graphical display by a statistical significance test. One could for instance think about *t*-tests or an analysis of variance to compare the phase means. Equivalently, a regression analysis can be performed to compare phase means for a specific student, including dummy variables indicating the phase:

$$Score_i = \beta_0 + \beta_1(phase2)_i + \beta_2(phase3)_i + e_i \qquad (1)$$

with $score_i$ equal to the percentage of accurate structural components in session *i*, and $phase2_i$ and $phase\ 3_i$ equal to 1 if session *i* is part of the second, respectively third phase, 0 otherwise. While the intercept, β_0, can be interpreted as the expected outcome in the baseline phase,

β_1 and β_2 reflect the effect of the teacher editing and the writer immersion, as compared to the baseline phase. The estimate of the intercept will be equal to the mean observed baseline score (for the first student, this is 30 %), the estimates of the regression coefficients of *phase2* and *phase3* as the difference in the observed means (40.25 % and 53.33 % for the first student). Testing the regression coefficients yields information about the evidence against the null hypothesis that there is no effect at all. For the first student, both effects appear to be statistically significant (p <.001).

Unfortunately, the results of an analysis of variance, *t*-test, or regression analysis assume that residuals (e_i) are independent, an assumption that is likely to be violated: there is probably some autocorrelation (Busk & Marascuilo, 1988). For instance, it is likely that subsequent residuals are more alike, due to time varying factors that are not controlled for in the model. Variables that have an influence at a specific moment will often also affect subsequent observations. The analyses further assume that scores are normally distributed, although they are relatively robust to violations of this assumption if groups are of comparable size, if they are not too small, and if the shapes of the distributions are similar (Posten, 1978). Techniques that were developed for the analysis of time series could be used, to account for the problem of autocorrelation, but these techniques require a large number (a minimum of 50 to 100) of measurements in each phase (Box & Jenkins, 1970; see Gorman & Allison, 1997, for an extensive discussion of statistical analyses of single-case data). In the following, we will see how the simple regression model can be extended to a hierarchical linear model to aggregate the results from several cases, that can easily be adapted to account for autocorrelation, even if for each case a small number of observations is available.

Aggregating single-case results

The graphs clearly suggest that at least for the cases participating in the study, there is a positive effect of the treatment on the quality of writing. The results suggest that there will

probably be an effect for similar cases. An appealing question now is whether we can indeed expect a positive effect for a new case, and how large this effect is. Furthermore, in view of generalizability, an answer on the question whether the effect varies over cases is called for. Especially if the effect appears to vary over cases, we may search for an explanation of this variation, by exploring whether the effect varies according to known characteristics of the cases. In the example, we could explore whether the difference in performance between the baseline phase and the writer immersion phase depends on the presence of an intermediate phase. In this paragraph, we combine and compare systematically the results of the seven cases by means of a hierarchical linear model, in order to answer these questions.

Hierarchical linear models were developed to analyze clustered data, for instance data stemming from pupils that can be grouped according to the school they belong to. In a hierarchical linear model, a regression equation is used to describe the variation of the scores within groups. In contrast to an ordinary regression equation the coefficients of the regression equation are allowed to vary over groups, and this variation is described by means of one or more additional regression equations (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). In our example, we have a similar structure: the scores we obtained can be grouped according to the pupil they stem from. We already saw a regression equation that can be used to describe the data of one case (Equation 1). For each case, we could define a similar regression equation. To write this set of seven equations in one single equation, we use an additional index *j* to indicate the student:

$$Score_{ij} = \beta_{0j} + \beta_{1j}(phase2)_{ij} + \beta_{2j}(phase3)_{ij} + e_{ij} \qquad (2)$$

$Score_{ij}$ now indicates the percentage of accurate structural components for student *j* in session *i*. The index *j* is also added to the regression coefficients, indicating that these coefficients can vary over cases. At the higher level of the hierarchy, the level of the cases, this variation of

regression coefficients is described with additional regression equations. The most basic equations state that the regression coefficients vary around a mean value:

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \\ \beta_{2j} &= \gamma_{20} + u_{2j} \end{aligned} \qquad (3)$$

It is assumed that the units at this level form a sample out of a population of units, and typically, that the regression coefficients of the first level (the β's) are normally distributed within this population. To that end, the residuals (*u*'s) are defined to follow a normal distribution with zero mean. The coefficient γ_{00} can now be interpreted as the expected performance (this is the population mean) under the baseline condition, γ_{10} and γ_{20} as the expected effect of teacher editing and writer immersion, respectively. u_{0j} indicates the degree to which the baseline performance from case *j* deviates from this expected performance, while u_{0j} and u_{0j} refer to the deviation of the effects of the treatments for case *j* as compared to the expected effects.

In the analysis, Equations 2 and 3 are regarded as one single model, and the γ's, as well as the variances of the residuals (σ_e^2, $\sigma_{u_0}^2$, $\sigma_{u_1}^2$, and $\sigma_{u_2}^2$), are estimated. It is also possible to assume that the different kinds of residuals at the second level covary, and to estimate the covariances ($\sigma_{u_0u_1}$, $\sigma_{u_0u_2}$, and $\sigma_{u_1u_2}$). Note that the residuals (the *e*'s and *u*'s), as well as the individual regression coefficients, the β's, are not estimated, although they could be estimated afterwards, as will be discussed further on. To estimate the parameters of the model by means of the commonly used restricted maximum likelihood procedure, we used the procedure MIXED from SAS (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006). Due to the small number of cases, especially in the second phase, we simplified the model by

excluding the covariances, and by excluding the variance of the effect of the second phase. Results are given in Table 1, in the column headed Model 1. The code for performing the analysis in SAS and in SPSS is given in the Appendix.

Insert Table 1 about here

It can be seen that the expected percentage of correct structural components is 39 % under the baseline condition, while it is respectively 33.02 % and 43.58 % higher in the teacher editing and writer immersion conditions. Note that the effect of the second condition is based on the data of the three cases of the first experiment only, since this phase was omitted in the second experiment. As a rule of thumb, if the estimate is twice as large as the standard error, the parameter differs significantly from zero when performing a two-sided test with a .05 significance level, since the ratio of the estimate over the standard error follows approximately a standard normal distribution (although it may be more accurate for the regression parameters to compare the ratio to a *t*-distribution, and to use a likelihood ratio test for testing the variance components; see Snijders & Bosker, 1999, for more details). Anyway, it is clear that the difference between baseline and both treatment conditions is highly significant. Comparing the estimate of the variance components with their corresponding standard errors reveals that there is no convincing evidence for differences between cases in the baseline performance, as well as in the effect of writer immersion. This is confirmed by performing a likelihood ratio test: the *p*-values are .08 and .48 respectively.

Although there is no compelling evidence for variation in baseline performance between cases, nor for any effect of writer immersion, we could test whether there are differences between the cases from the first and the second experiment, by extending the regression equations describing the variation over cases, including the group as an additional independent variable:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}(group2)_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} \\ \beta_{2j} &= \gamma_{20} + \gamma_{21}(group2)_j + u_{2j}\end{aligned} \quad (4)$$

where $(group2)_j$ equals 1 if case *j* belongs to the second group, 0 otherwise. Now γ_{00} can be interpreted as the expected performance in the baseline condition for a case of the first group, γ_{01} as the increase in this expected value if the case belongs to the second group. The expected effect in the first experiment is indicated by γ_{20}, the additional effect in the second experiment by γ_{21}. Results of this model are also presented in Table 1 (Model 2). The expected performance in the baseline condition is 33.63 % for the first group, while it is 9.57 % higher for the second experiment, a difference that is, however, not statistically significant, $t = 1.28$, $df=59$, $p = .21$. In the first experiment, cases score 54.23 % higher in the writer immersion condition, while in the second experiment the effect of writer immersion is 21.98 % smaller (and so equals only 32.25%). The difference between the baseline phase and the writer immersion phase thus is larger in the first experiment than in the second one, a difference that is statistically significant $t = -2.88$, $df = 59$, $p = .006$. Further note that residual differences between cases again are not statistically significant. The estimated variance of the between-cases variance of the effect of writer immersion even equals zero. It may look strange that the estimate is exactly zero. This is because, after taking the group into account, observed differences between cases are even smaller than could be expected based on chance alone. This situation would lead to a negative estimate of the variance (see Snijders & Bosker, 1999), which is of course outside the range of possible values, and therefore the estimate is set to zero. This small variance estimate is no surprise: in the preceding model, we have already seen that there are only small differences between cases. Since there seems to be a substantial difference between both groups of cases, differences between cases after accounting for the group, must indeed be very small.

Where we described a regression analysis for one case, we discussed two problems: the small number of observations, as well as the possible autocorrelation in residuals. One merit of using hierarchical linear models for aggregating single-case results is that –since the regression coefficients are not estimated for each case separately, but conclusions rather focus on the population of cases- a small number of observations per case is allowed. Cases with only a very few measurements (or even only one) also yield (little) information about the population of cases. More important for the analysis is the number of cases, which is in our example rather small, a problem that will be discussed later on.

Further, the hierarchical linear models we described before still require independent residuals, but the models easily can be adapted to accommodate to this problem. To explore a possible autocorrelation, we re-estimated the parameters of the second model, extended with a first-order autocorrelation parameter (see Appendix for the SAS and SPSS code). The estimate of the autocorrelation parameter appeared to be relatively small and statistically not significant (p = .65 when using a likelihood ratio test), and therefore all other parameter estimates and standard errors remain approximately the same (Table 1, Model 3), and conclusions are unaffected.

Discussion and conclusions

The use of hierarchical linear models allows summarizing the findings of several cases examined in the same or in several studies in a systematic and quantitative way. More important, by aggregating single-case results, conclusions are not necessarily restricted to the studied cases, but may refer to a broader population as well. The use of hierarchical linear models permits, for instance, assessing the effect of a certain treatment for ‘an average case’, or, otherwise stated, the expected effect if a new case from the population were investigated. They further give the opportunity to evaluate the degree to which the effect varies over cases,

and to look for characteristics of persons, settings, or treatments that have an influence on the effect. By aggregating the results of several cases, we increase the power for assessing an overall effect of a treatment. It is, for instance, possible that for a set of cases, no significant treatment effect is found although there is a tendency for an effect in a certain direction. When combining the results of the cases, the small parts of evidence are accumulated, and the overall treatment effect may become visible (i.e., statistically significant), and the size of the effect can be estimated accurately.

Results of a set of cases thus combines the merits of single-case designs (giving information about individual cases) and of group designs (giving information about an average case). Yet, the aggregation of single-case results may also lead to improved estimates for the effects for individual cases. Based on the results of the analysis using hierarchical linear models, one could calculate estimates for the regression coefficients for specific cases (the β's from Equation 2), that are an optimally weighted mean of the overall regression coefficients (the γ's from Equations 3 and 4) and the regression coefficients that would have been obtained when performing an ordinary regression analysis for the separate cases (Equation 1). Especially when the number of observations for a case is small, and therefore the regression coefficients that would be obtained if using the data from that case only are relatively unreliable, and when the effect does not appear to vary largely over cases, the profit of using these estimates, called empirical Bayes estimates, is large (Snijders & Bosker, 1999).

In the example, only relatively simple hierarchical linear models are illustrated. Hierarchical linear models however can easily be adapted, for instance to aggregate more complex phase designs, such as withdrawal or interaction designs, or alternation designs such as completely randomized single-case designs. The model could also include a second independent variable, yielding a factorial single-case design that can be used to evaluate the main effects of both

variables, as well as their interaction effect (see Barlow & Hersen, 1984, and Edgington & Onghena, 2007, for more information about these and other single-case designs). A time-varying covariate, such as the session number, can be included in the model describing the individual scores (Equation 2) in order to model growth over time for instance due to maturation. An interaction term between this covariate and the phase indicator is included to model a possible differential growth in the conditions. Additional regression coefficients (e.g., the linear trend) may be defined as varying over cases, and more complex variance structures for the residuals can be defined. For instance, while in Equation 2, it is assumed that the residual variance is the same in the three phases, this restriction can be dropped and separate variance components for each phase can be defined. Finally, these hierarchical linear models can further be adapted to aggregate single-case results that are summarized by means of measures of effect size. Several extensions are discussed and illustrated by Van den Noortgate and Onghena (2003a; 2003b).

Although the aggregation of single-case data using hierarchical linear models has a lot of potential assets, aggregation should be performed with care, and may even turn out difficult or impossible for certain designs or research settings. A first consideration is that usually in single-case research, cases are not sampled randomly. Analyses performed or conclusions drawn in single-case research usually also do not require random sampling, since the focus is on validly assessing the effect for the case that is concerned, not to generalize the results to a broader population. Yet, the assumption that cases form a random sample out of a population of cases must be made when generalizing the results based on a set of single-case data. The researcher therefore should carefully examine and describe the characteristics of the cases involved. In the study that was used for the example, participants were also purposefully selected: cases were persons with behavioral and learning disabilities “chosen because of many structural errors in their writing, as well as their ability to write functionally” (Lawson

& Greer, 2006, p. 153). It is clear that results are not to be generalized to cases that differ from the cases selected for the study. Note that this problem also often occurs in group experimental research, since group experiments are also frequently carried out on a sample that was not (completely) randomly drawn (Edgington & Onghena, 2007).

A second limitation is that it may be difficult to model all the important patterns shown by the data. For instance, in the example, we merely compared the mean performance in the three phases (as did the original authors, Lawson & Greer, 2006). Yet, the graphs suggest that the effect is not immediately fully present, which is not surprising since during the intervention phase, students learn from the feedback they are given in each session. Furthermore, in the writer immersion phase, ceiling effects are likely to be present. This suggests that the effect that is estimated is probably an underestimation of the real effect. These phenomena –and others such as cross-over effects– can in principle be modeled, but the modifications would make the model much more complex and therefore more difficult to estimate and interpret. Whenever possible, graphical displays can (and, in many instances, should) be used to qualify and supplement the results of the hierarchical linear analysis.

Another point of concern is the number of cases included in the analysis. It is clear that with only a small number of cases available, it is not possible to get reliable estimates of population characteristics, such as the mean effect and variation over cases of this effect. Moreover, it will be difficult to assess an existing overall effect or possible moderator effects, unless they are as large as seems to be the case in the example. The number of cases is of concern for a second reason: the parameters of hierarchical linear models are commonly estimated using maximum likelihood procedures, and tests are based on large sample properties of maximum likelihood estimates. This means that if the results of only a few cases are aggregated, as in the example, results should be considered as only indicative. To perform the analysis comfortably, at least about 20 cases are recommended, or even more in case

several independent variables are included and/or several (co)variance components have to be estimated at the level of cases.

The issue of the number of cases may be even more problematic in case hierarchical linear models are used to combine the single-case results of several studies. Single-case studies are often very heterogeneous regarding procedures, situations, and subject characteristics (Salzberg, Strain, & Baer, 1987). Heterogeneity could be regarded as a source of information since it allows searching for variables that moderate the effect. As illustrated in the example, those variables could be included as independent variables in the hierarchical linear model. Yet, if studies and cases differ with respect to a large number factors, modeling a possible moderator effect of these factors would imply a considerable extension of the model at the level of cases, requiring a substantial number of cases.

Finally, we note that if the treatment is randomly assigned to measurement occasions or if the phase change is randomly chosen, a nonparametric randomization test can be used to evaluate the treatment effect (Edgington & Onghena, 2007; Ferron & Onghena, 1996; Todman & Dugard, 2001). Also for aggregating single-case results from simultaneous or sequential replication designs, randomization tests or other nonparametric procedures have been proposed (Edgington & Onghena, 2007; Ferron & Sentovich, 2002). The attractiveness of the tests lies in the fact that they do not resort to distributional assumptions and are applicable even for small numbers of observations and cases. Yet, the hierarchical linear models approach offers more possibilities, including a systematic search for moderator variables, or estimating the size of the treatment and moderator effects and as well as the variation between cases in these effects.

To conclude, we want to stimulate behavioral researchers and practitioners to continue replicating single-case studies and to consider using hierarchical linear models to aggregate

single-case results of their own studies and/or from the ones from others, allowing to make inferences at the level of the individual cases, as well as at the group level.

References

Barlow, D. H., & Hersen, M. (1984). *Single-case experimental designs: Strategies for studying behavior change* (2nd ed.). New York: Pergamon Press.

Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis, forecasting and control*. San Francisco: Holden-Day.

Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment, 10*, 229-242.

Caramazza, A. (1990). *Cognitive neuropsychology and neurolinguistics: Advances in models of cognitive function and impairment*. Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cook, D. J. (1996). Randomized trials in single subjects: The N of 1 study. *Psychopharmacology Bulletin, 32*, 363-367.

Edgington, E. S., & Onghena (2007). *Randomization tests* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.

Edgington, E. S. (1996). Randomized single-subject experimental designs. *Behaviour Research and Therapy, 34*, 567-574.

Ferron, J., & Onghena, P. (1996). The power of randomization tests for single-case phase designs. *Journal of Experimental Education, 64*, 231-239.

Ferron, J., & Scott, H. (2005). Multiple baseline designs. In B. S. Everittt, & D. C. Howell (Eds), *Encyclopedia of statistics in behavioral science* (Vol. 3 pp. 1306-1309). Chichester, UK: John Wiley & Sons.

Ferron, J., & Sentovich, C. (2002). Statistical power of randomization tests used with multiple-baseline designs. *Journal of Experimental Education, 70*, 165-178.

Franklin, R. D., Allison, D. B., & Gorman, B. S. (1997). Introduction. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 1-12). Mahwah, NJ: Lawrence Erlbaum.

Gorman, B. S., & Allison, D. B. (1997). Statistical alternatives for single-case designs. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 159-214). Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Ittenbach, R. F., & Lawhead, W. F. (1997). Historical and philosophical foundations of single-case research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 13-39). Mahwah, NJ: Lawrence Erlbaum.

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York: Oxford University Press.

Kratochwill, T. R., & Levin, J. R. (1992). *Single-case research design and analysis. New directions for Psychology and Education*. Hillsdale, NJ: Lawrence Erlbaum.

Lawson, T. R., & Greer, R. D. (2006). Teaching the function of writing to middle school students with academic delays. *Journal of Early and Intensive Behavior Intervention, 3*, 151-170.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS® system for mixed models* (2nd ed.). Cary, NC: SAS Institute Inc.

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23*, 341-351.

Onghena, P. (2005). Single-case designs. In B. S. Everittt, & D. C. Howell (Eds), *Encyclopedia of statistics in behavioral science* (Vol. 3 pp. 1850-1854). Chichester, UK: John Wiley & Sons.

Pillemer, D. B. (1984). Conceptual issues in research synthesis. *Journal of Special Education, 18*, 27-40.

Posten, H. O. (1978). The robustness of the two-sample t test over the Pearson system. *Journal of Statistical Computation and Simulation, 6*, 295-311.

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

Salzberg, C. L., Strain, P. S., & Baer, D. M. (1987). Meta-analysis for single-subject research: When does it clarify, when does it obscure? *Remedial and Special Education, 8*, 43-48.

SAS Institute. (2004). *SAS/STAT 9.1 User's guide*. Cary, NC: SAS Publishing.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

SPSS Inc. (2005) *Linear mixed modeling in SPSS* [Web Page]. URL http://www.spss.com/home_page/wp127.htm [2007, January 17].

Todman, J. B., & Dugard, P. (2001). *Single-case and small-n experimentel designs: A practical guide to randomization tests*. Mahwah: Erlbaum.

Van den Noortgate, W., & Onghena, P. (2003). Combining single-case experimental studies using hierarchical linear models. *School Psychology Quarterly, 18*, 325-346.

Van den Noortgate, W., & Onghena, P. (2003). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, & Computers, 35*, 1-10.

Appendix: SAS and SPSS codes for the models from the example

SAS

The dataset, named *Lawson*, consists of one row for each score on the dependent variable. The number of rows per case therefore is equal to the number of sessions during which the case was measured. Besides the dependent variable, called *Percent*, it includes indicator variables for the case and session (called *Case* and *Session*), for the second group (*Group2*), and the teacher feedback and writer immersion treatment (*Phase2* and *Phase3*). For information about managing data sets, we refer to the Help menu, as well as to SAS Institute (2004). The first model is fitted by running the following code:

```
PROC MIXED DATA=Lawson COVTEST;
CLASS Case;
MODEL Percent= Phase2 Phase3 / SOLUTION;
RANDOM Intercept Phase3 / SUB= Case;
RUN;
```

In the first statement, the MIXED procedure is called. The DATA= statement refers to the data set in which the data are stored. The COVTEST option is added to get standard errors and approximate *p*-values for the variance components. In the second statement, the indicator variable *Case*, is defined as a categorical variable. In the third line, the fixed part of the model is described. The variable *Percent* is defined as the dependent variable, the variables *Phase2* and *Phase3* as independent variables. The model includes an intercept by default. The SOLUTION-option is used to request in the output the estimates, standard errors, *t*-statistics and *p*-values for significance testing for all fixed effects

The RANDOM statement is used to describe the random part of the model. We indicate that the intercept as well as the coefficient of *Phase3* can vary randomly across cases. The code is closed by the RUN-command.

For the second analysis, the model is extended by including additional independent variables. This is done by extending the MODEL-statement as follows:

MODEL *Percent= Phase2 Phase3 Group2 Group2*Phase3* / SOLUTION;

Finally, to estimate the first-order autocorrelation, a single statement is added (before the RUN-command), while the rest of the code remains unchanged:

REPEATED / TYPE=AR(1) SUB= *Case*;

While in the RANDOM statement, the random part on the second level is described, in the REPEATED statement the same is done for the first level. The option TYPE=AR(1) requests modeling a first-order autocorrelation within cases.

SPSS

The codes in SPSS are highly similar. Extensive information is found in the Help menu, as well as in SPSS (2005). For the first model, we run the following code:

MIXED *Percent* WITH *Phase2 Phase3*
/PRINT = SOLUTION TESTCOV
/FIXED = *Phase2 Phase3*
/RANDOM *Intercept Phase3* | SUBJECT(*Case*).

For the second model, we have:

MIXED *Percent* WITH *Phase2 Phase3 Group2*
/PRINT = SOLUTION TESTCOV
/FIXED = *Phase2 Phase3 Group2 Group2*Phase3*
/RANDOM *Intercept Phase3* | SUBJECT(*Case*).

To model a possible first-order autocorrelation, we write:

MIXED *Percent* WITH *Phase2 Phase3 Group2*
/PRINT = SOLUTION TESTCOV
/FIXED = *Phase2 Phase3 Group2 Group2*Phase3*

/RANDOM *Intercept Phase3* | SUBJECT(*Case*)

/REPEATED | COVTYPE(AR1) SUBJECT(*Case*).