Hierarchical linear models for the quantitative integration of effect sizes in single-case research

WIM VAN DEN NOORTGATE and PATRICK ONGHENA
Katholieke Universiteit Leuven, Leuven, Belgium

In this article, the calculation of effect size measures in single-case research and the use of hierarchical linear models for combining these measures are discussed. Special attention is given to meta-analyses that take into account a possible linear trend in the data. We show that effect size measures that have been proposed for this situation appear to be systematically affected by the duration of the experiment and fail to distinguish between effects on level and slope. To avoid these flaws, we propose to perform a multivariate meta-analysis on the standardized ordinary least squares regression coefficients from the study-specific regression equations describing the response variable.

While in the last decades both single-case experimental designs and meta-analysis or quantitative research synthesis have become increasingly popular in several research domains, the meta-analysis of single-case studies has received relatively little attention, and relatively few meta-analyses of single-case data have been published (Busk & Serlin, 1992; Busse, Kratochwill, & Elliott, 1995).

Nevertheless, we believe that integrating the results of single-case studies can yield important supplementary information that would not be available otherwise. Although the results of a single-case experiment are especially informative with respect to the effectiveness of a specific treatment for the case investigated, a meta-analysis of single-case studies gives information about the overall effect, without losing information about specific cases. A meta-analysis also allows one to establish whether the effect varies from case to case and to investigate a possible moderating influence of case or study characteristics on the treatment effect. In contrast to traditional narrative reviews of single-case studies, a meta-analysis approaches all these questions in a quantitative and systematic way.

Performing a meta-analysis often means combining standardized measures of effect size, rather than combining raw data. Indeed, often, not all raw data are available to a meta-analyst. Moreover, in different studies, the dependent variable may be measured on different scales. In the article, we discuss possible measures of effect size that take into account the peculiarities of single-case data and the use of hierarchical linear models for the meta-analysis of these effect sizes. Our purpose is to stimulate behav-

ioral researchers to consider this powerful and general data-analytic strategy.

To introduce the use of hierarchical linear models in single-case research for the meta-analysis of measures of effect size, we will start with a simple situation in which a meta-analyst wants to combine the results of single-case studies with AB phase designs, without trends in the data. Next, we will discuss meta-analyses that take possible within-phase linear trends into account. We will discuss proposed measures for effect size and will present an alternative approach, illustrated with an example. Finally, we will treat the meta-analysis for more complex single-case designs and will discuss extensions and alternatives for the approach and models presented—for instance, to account for a possible serial dependency.

The Meta-Analysis for Single-Case AB Phase Designs

A popular measure by which to express the magnitude of an effect in a group experimental design is the following measure proposed by Cohen (1969):

$$d = \frac{\bar{X}_E - \bar{X}_C}{SD}. \quad (1)$$

This is the difference between the mean score in the experimental group and the mean score in the control group, divided by a standard deviation. The standard deviation used is the estimate of the standard deviation of the control condition or, assuming that the population standard deviations are equal, the pooled estimate of this common standard deviation.

Hedges (1981) showed that if the scores are independently normally distributed and if a common population variance is estimated by pooling the two sample estimates, the sampling distribution of d is linearly related to a noncentral t distribution but is closely approximated by

Correspondence concerning this article should be addressed to W. Van den Noortgate, Department of Education, Katholieke Universiteit Leuven, Vesaliusstraat 2, B-3000 Leuven, Belgium (e-mail: wim.vandennoortgate@ped.kuleuven.ac.be).

a normal distribution with a mean equal to the population effect size and with a standard deviation estimated by

$$\hat{\sigma}(d) = \sqrt{\frac{n_C + n_E}{n_C n_E} + \frac{d^2}{2(n_C + n_E)}}, \quad (2)$$

with n_C and n_E being the sizes of the control and the experimental groups.

Assuming that the scores in a single-case experiment with a baseline and an intervention phase are independently normally distributed with a common variance, we can use an analogous measure, the difference between the phase means divided by the pooled within-phase standard deviation, and can use the meta-analytic techniques of Hedges and Olkin (1985) or others for the meta-analysis of the results of several single-case studies, although they were initially developed for the meta-analysis of group comparison studies (Gingerich, 1984).

Thanks to the tradition of visual inspection of data in single-case research, results are often presented in time series graphs from which raw data can be retrieved and used for the calculation of d . From time to time, the raw data from single-case studies cannot be retrieved, but data or effects are reported by summary statistics, test statistics, p values, or measures of effect. Again, these can often be used to calculate d , on the basis of conversion formulas available in the meta-analytic literature (e.g., Rosenthal, 1994).

Bryk and Raudenbush (1992) described how an adaptation of the general hierarchical linear model can be used for a meta-analysis of measures of effect or, generally, for hierarchically structured data sets in which the results of each study are summarized in a descriptive statistic with an approximately normal sampling distribution with a known (or already estimated) variance. Bryk and Raudenbush call such models *variance-known models*.

The simplest hierarchical linear model for combining the effect sizes of single-case research is a two-level model: Measurements are grouped in cases. On the first level (the level of measurements), we have the following model:

$$d_j = \delta_j + e_j, \quad (3)$$

with d_j being the observed effect size of case j , δ_j the “true” effect size of case j , and e_j the sampling error for the observed effect size from case j . On the second level (the level of cases), we can write the true effect size of case j as deviating from an overall effect size:

$$\delta_j = \gamma_0 + u_j, \quad (4)$$

with γ_0 being the overall effect size and u_j a deviation typical of case j . Combining both models, we get

$$d_j = \gamma_0 + u_j + e_j. \quad (5)$$

If in at least some studies, more than one case is included, the model can be extended by using an additional level. The model thus becomes a three-level model: Measurements are grouped in cases, which in turn are grouped in studies:

$$d_{jk} = \gamma_0 + v_k + u_{jk} + e_{jk}, \quad (6)$$

with γ_0 the overall effect size, v_k a deviation typical of study k , u_{jk} a deviation typical of case j from study k , and e_{jk} the sampling error for the observed effect size for case j from study k .

Another extension is the inclusion of study or case characteristics as covariates on the study or case level in order to explain a possible variation between studies or cases. In a three-level model, the moderating effect of a case characteristic can be defined as varying over studies by modeling an additional residual term. As an example, a three-level hierarchical linear model including a case and a study characteristic as predictors and with a study-specific influence of the case characteristic is presented in Equation 7:

$$d_{jk} = \gamma_0 + (\gamma_1 + v_{1k})X_{jk} + \gamma_2 Z_k + v_{0k} + u_{jk} + e_{jk}, \quad (7)$$

with γ_0 the expected effect size for X and Z equal to zero, X_{jk} the value for case j from study k on the case characteristic X , γ_1 the overall moderating effect of the case characteristic X , v_{1k} the study-specific deviation from the overall effect of the case characteristic X , Z_k the value for study k on the study characteristic Z , and γ_2 the effect of the study characteristic Z .

Residuals on each level are usually assumed (multivariate) normally distributed with means equal to zero. The parameters of these hierarchical linear models can be estimated by using specific user-friendly software for multi-level models (for instance, MLwiN [Goldstein et al., 1998] and HLM [Bryk, Raudenbush, Congdon, & Seltzer, 1988]) or by using a general statistical package (for instance, SAS [see Sheu & Suzuki, 2001; Wang & Bushman, 1999] or S-Plus [see Pinheiro & Bates, 2000]), assuming normal distributions for the residual terms on each level.

The Meta-Analysis for Single-Case AB Phase Data Containing Trends

Sometimes, the response of a case decreases or increases systematically over time. The standardized mean difference may be affected by such a trend. For instance, in a case in which there is no effect of the intervention, a positive linear trend in the data results in a higher expected score in the second phase, as compared with the first phase, and thus in a positive standardized mean difference.

Several regression techniques have been proposed to estimate the effect size for a specific case by taking a trend into account. Major approaches are the piecewise regression approach of Center, Skiba, and Casey (1985–1986), the initial baseline-only regression approach of Allison and Gorman (1993), and the approach of White, Rusch, Kazdin, and Hartmann (1989). In the next section, we first will discuss these three approaches and propose an alternative approach for describing the effect. Next, we will discuss the meta-analysis for the presented measures of effect size, illustrated with an example.

Effect size measures for single-case data showing trends. Center et al. (1985–1986) suggested measuring the effect of a certain intervention for a specific case on the basis of a linear regression model including a phase indi-

cator, a time related variable t (for instance, age or duration), and possibly, the interaction between both:

$$Y_i = \beta_0 + \beta_1(\text{phase})_i + \beta_2 t_i + \beta_3(\text{phase})_i t_i + e_i, \quad (8)$$

with $(\text{phase})_i$ equaling 1 if score i is observed during the intervention phase and 0 otherwise.

β_0 indicates the baseline level at time $t = 0$, and β_2 indicates the linear trend in the baseline scores. β_1 can be interpreted as the benefit of the intervention at time $t = 0$, whereas β_3 indicates the effect of the intervention on the trend. The interpretations of β_0 and β_1 from Equation 8 are often theoretically not interesting, because $t = 0$ is often situated before the baseline phase or even cannot occur, by definition. Therefore, t from Equation 8 is sometimes replaced by t^* , which is defined as the value of t minus the value of t at the last observation of the baseline phase (Center et al., 1985–1986) or at the first observation of the intervention phase (Huitema & McKean, 2000). For instance, if the variable t is equal to the measurement number, t^* would be equal to $(t - n_A)$ or $t - (n_A + 1)$, respectively, where n_A represents the number of observations in the baseline phase. β_1 thus indicates the immediate effect of the intervention at the start of the treatment phase. Although β_1 may be affected by the choice for one of both definitions of t^* , this influence is substantial only if the difference in slopes is large, in case β_1 is often of minor importance as compared with β_3 .

To express the level change, Center et al. (1985–1986) proposed to constrain the coefficient β_1 to zero and to compare the variances explained by the full and the reduced model. In a first step, an F ratio is calculated, using the following equation:

$$F = \frac{MS(\text{effect})}{MS(\text{error})}, \quad (9)$$

with $MS(\text{effect})$ being the difference between the error sum of squares from the full and the reduced model, divided by the number of terms eliminated and $MS(\text{error})$ the error sum of squares of the full model divided by the number of observations minus the number of parameters in the full model.

In a second step, the F ratio can be converted to the standardized mean difference, using the following formula (Rosenthal, 1994):

$$d = 2 \sqrt{\frac{F}{df_{(\text{error})}}}. \quad (10)$$

Similarly, the effect of the intervention on the slope (trend) or a combination of both effects can be expressed by the d statistic. Hence, the procedure results in three different measures of effect size, indicating the effect on the level, the effect on the slope, and the total effect. Because it is not always clear which one to use or how to use several measures at one time, the procedure has not been generally accepted (Busse et al., 1995).

Allison and Gorman (1993) argued that in the procedure of Center et al. (1985–1986), constraining β_1 and/or β_3 in Equation 8 to zero often leads to an underestimation of the effect of the intervention. A part of the variance ex-

plained by these factors indeed moves to the variance explained by the trend, rather than to the residual variance. Therefore, Allison and Gorman (1993) proposed to use the baseline data only to estimate the trend and to correct the scores of both phases for the expected scores owing to this trend. These residual scores are then regressed on both the phase and the phase time interaction:

$$(Y_i - \hat{Y}_i) = \beta_0 + \beta_1(\text{phase})_i + \beta_3(\text{phase})_i t_i^* + e_i. \quad (11)$$

The F ratio, which is obtained by dividing the mean square regression by the mean square residual, is then transformed to a d statistic, and an appropriate sign is given.

An alternative measure that combines the effect of the intervention on both the slope and the level was suggested by White et al. (1989). They proposed to calculate a regression equation for each phase and to compare the expected scores at the end of the experiment. A corrected standardized mean difference d can be calculated:

$$d = \frac{\bar{Y}_B^* - \bar{Y}_A^*}{SD^*}, \quad (12)$$

with $\bar{Y}_B$ being the predicted score on the last occasion of the treatment phase, based on the treatment phase regression equation, $\bar{Y}_B^*$ the predicted score on the last occasion of the treatment phase, based on the baseline phase regression equation, and SD^* the within-phase standard deviation around the regression lines. Instead of linear regression equations, it is possible that nonlinear time functions can be used to make predictions (Hershberger, Wallace, Green, & Marquis, 1999). Moreover, the point in time for which the score is predicted may be another point that is of interest scientifically (Hershberger et al., 1999).

However, as we will demonstrate next, both the procedure of Allison and Gorman (1993) and that of White et al. (1989) are limited in describing the effect of an intervention. Two basic shortcomings can be identified. First, both kinds of effect size estimates depend on the length of the experiment. Second, different effects can lead to equal estimates of the effect size.

Suppose Figure 1 represents the effects of an intervention for two cases. In the baseline phase, scores of both cases are approximated by a regression line with an intercept of 1.6 and a slope of 1. For the first case, there is no immediate effect on the level of performance, but there is an effect on the linear trend (the slope is doubled). For the second case, there is no effect on the linear trend, but scores are almost 14 points higher than would have been expected on the basis of the baseline regression line.

From the figure, it can be seen that the length of the experiment can indeed influence the expected effect size estimates of Allison and Gorman (1993). If the two cases were not measured on the last three occasions, for both cases the residual variance would have been approximately the same, but the total variance would have been reduced (especially for Case 1). Therefore, the effect size estimates would differ from those obtained after the 20th occasion (Table 1).

In the procedure of White et al. (1989), the estimate of the effect size also depends on the length of the experi-

Figure 1. The dependence of the measure of effect of Allison and Gorman (1993) and White, Rush, Kazdin, and Hartmann (1989) on the length of the experiment.

ment, in cases in which there is an effect of the intervention on the trend. In Figure 1, we see that if the treatment has a positive effect on the level and slope (Case 1), one can expect a larger effect size estimate if an experiment stops after 20 measurement occasions, rather than at the 17th occasion. For Case 2, in which there is no effect on the slope, the effect size measure, however, is not systematically affected by the duration of the experiment (Table 1). The arrows drawn in the figure indeed indicate that if the effect is evaluated at Occasion 20 instead of at Occasion 17, the distance between the regression line based on the baseline scores and the regression line based on the scores from the treatment phase gets larger for Case 1 but remains the same for Case 2.

In addition, the example reveals that in both procedures, different effects can lead to equal effect sizes. Although both baselines are equal, the effect of the intervention is different in both cases. The effect of the intervention is initially larger for Case 2, but after a while, Case 1 seems to benefit most from the intervention. However, if residuals (for all 20 occasions) calculated using the baseline regression line are regressed on the phase $\times$ time phase interaction (Equation 8), the mean square regression and the mean square residual are approximately equal for both cases (467.50 and 10.16, respectively), giving equal F s and equal effect size estimates in the Allison and Gorman

procedure (1993) (Table 1). Figure 1 also illustrates that the White et al. (1989) effect size estimates for both cases are very similar if evaluated after the 17th observation.

Although it is true that additional data can make an observed effect more impressive, this is a matter that should be reflected in a statistical test or in a confidence interval of the effect size measure, not in the effect size measure itself. In order to describe the effect of an intervention, independently of the length of the experiment and without losing important information about the kind of effect, one single measure of effect size cannot suffice. Therefore, we propose an alternative procedure. Like Center et al. (1985–1986), we suggest using a regression equation in which the scores are regressed on the phase, the trend, and the trend $\times$ phase interaction and calculating the regression coefficients on the basis of the ordinary least squares (OLS) criterion. However, instead of summarizing the effects by calculating the effect size on the basis of the F ratio, we suggest using the regression coefficients of this regression equation as measures of the effect. By using both coefficients β_1 and β_3 from Equation 11 to describe the effect, both shortcomings described above can be avoided.

Unfortunately, the regression coefficients are not necessarily comparable over studies. First, the time-related variable might be measured on dissimilar scales. In different studies, this variable can, for instance, be defined as

Table 1
Calculation of the Effect Sizes for the Two Fictitious Cases of Figure 1

	Allison and Gorman (1993)	White, Rush, Kazdin, and Hartmann (1989)	Standardized Regression Coefficients
Case 1			
Evaluated after Observation 20	$F(2,17) = 46.00, d = 4.65$	$d = 6.09$	$\hat{\beta}_1^* = 0/\hat{\beta}_3^* = 0.61$
Evaluated after Observation 17	$F(2,14) = 16.29, d = 3.05$	$d = 3.92$	$\hat{\beta}_1^* = 0.20/\hat{\beta}_3^* = 0.53$
Case 2			
Evaluated after Observation 20	$F(2,17) = 46.00, d = 4.65$	$d = 4.16$	$\hat{\beta}_1^* = 4.16/\hat{\beta}_3^* = 0$
Evaluated after Observation 17	$F(2,14) = 35.83, d = 4.52$	$d = 3.83$	$\hat{\beta}_1^* = 4.27/\hat{\beta}_3^* = -0.06$

the age in months, as the day number since the start of the experiment, or as the observation number. To make coefficients comparable, one must convert this time variable to the same scale. Second, the dependent variable can be measured on different scales. Often, however, it is realistic to assume that the scales of the dependent variable are linearly equatable (Hedges & Olkin, 1985), in which case the regression coefficients for the covariates can be made comparable over studies by dividing them by the root-mean squared error. Consequently, the corresponding (co)variance matrix of the estimates must be divided by the mean squared error. Note that if there is no linear trend, there is only one covariate, the phase indicator. Dividing the corresponding regression coefficient β_1 by the root-mean squared error results in the standardized mean difference d . The use of the effect size measures we present thus simplifies to the use of the standardized mean difference if there are no linear trends.

The standardized regression coefficients for the data from Figure 1 are also given in Table 1. It can be seen that these standardized regression coefficients are affected to only a very small degree by evaluating the effect after 17 observations, instead of after 20 observations. This influence of the length of the experiment is due only to the random fluctuation of scores around the regression lines. Moreover, it can be seen that using the standardized regression coefficients allows a more complete description of the effects.

The Meta-Analysis of Standardized Regression Coefficients

Because the standardized regression coefficients are comparable over studies, one can use them for a meta-analysis of the results of several studies. Since there are two measures for the effect in each study, we suggest performing a multivariate meta-analysis. Kalaian and Raudenbush (1996) presented a multivariate hierarchical linear model that can be used to perform a multivariate meta-analysis.

On the first level, we have

$$\hat{\beta}_{1j} = \beta_{1j} + e_{1j} \text{ and } \hat{\beta}_{3j} = \beta_{3j} + e_{3j}, \quad (13)$$

with $\hat{\beta}_{1j}$ and $\hat{\beta}_{3j}$ equal to the standardized OLS estimates of the regression coefficients for level change and slope change for case j and β_{1j} and β_{3j} equal to the "true" standardized regression coefficients for level change and slope change for case j . On the second level,

$$\beta_{1j} = \gamma_1 + u_{1j} \text{ and } \beta_{3j} = \gamma_3 + u_{3j}. \quad (14)$$

On each level, residual terms are assumed to be multivariate normally distributed. Moreover, as is the case in a univariate meta-analysis, it is assumed that the Level 1 variance matrix is known. In practice, we use an estimate. This (estimated) covariance matrix of the regression coefficients can usually be retrieved when a linear regression analysis is performed with statistical software. Parameters can again be estimated with specialized hierarchical linear model software or by a general statistical package. For de-

tails about the estimation of the unknown parameters, we refer to Kalaian and Raudenbush (1996) and to the software manuals.

To perform a multivariate meta-analysis, we do not need an estimate of all regression coefficients for all studies. Thus, the approach can still be used when, for one or more studies, data are analyzed using the regression model of Center et al. (1985–1986) but only the effect on the performance level or on the trend is reported. The model can again easily be extended to include study or case characteristics as covariates on the second level or to include other levels (for instance, separate case and study levels).

An example. Marlowe, Madsen, Bowen, Reardon, and Logue (1978) designed an experiment to determine the relative effectiveness of teacher and counseling approaches in the reduction of disruptive or inappropriate classroom behavior. Inappropriate classroom behavior frequencies of 12 academically low-achieving, seventh-grade, black male students with a reported high rate of inappropriate classroom behavior were recorded. To evaluate counseling approaches, three groups of 4 students were randomly assigned to one of three treatment conditions: behavioral counseling, client-centered counseling, or no counseling.

In Figure 2, time series graphs are shown, representing the frequencies of off-task behavior in the baseline phase and the intervention phase. During the second phase, eight 30-min counseling sessions were conducted for the behavioral (cases from first column) and client-centered (second column) counseling groups, while the no-contact control group (third column) continued baseline without treatment.

We used the approach we presented above to combine the results from the time series graphs. Note that the cases differed from each other in the number of measurements in both phases, owing to missing values. After reconstructing the raw scores from the graphs, we performed separate ordinary regression analyses for each case, with the time variable defined as the day of observation minus the number of days in the first phase (10). The OLS regression coefficients, together with their covariance matrices, were standardized by dividing them by the root *MSE* and the *MSE*, respectively, and were used to perform a multivariate meta-analysis following the techniques of Kalaian and Raudenbush (1996). Calculations were done with the *MLwiN* software (Goldstein et al., 1998), assuming normal distributions for the residual terms. As was proposed by Kalaian and Raudenbush, the within-case residuals (e_{1j} and e_{3j}) were first made orthogonal with unit variance, using the Cholesky factorization (using the *CHOL* command in *MLwiN*). Setting up a multivariate model in *MLwiN* furthermore requires defining an additional lower level, without a Level 1 variation being specified, which is used solely to define the multivariate structure, using the dummies as predictors. We therefore used a three-level model, with the first level indicating the effect size measure, the second level the within-case level, and the third level the case level. The calculation of maximum likelihood estimates of the unknown parameters was done with the iterative generalized least

Figure 2. Time series graphs, representing the frequency of individual off-task behavior by day and condition, for the behavioral counseling group (first column), the client-centered counseling group (second column), and the no-contact group (third column). From "Severe Classroom Behavior Problems: Teachers or Counselors," by R. H. Marlowe, C. H. Madsen, Jr., C. E. Bowen, R. C. Reardon, and P. E. Logue, 1978, *Journal of Applied Behavior Analysis*, 11, 62-63. Copyright 1978 by the Society for the Experimental Analysis of Behavior. Reprinted with permission.

squares algorithm of the program. The parameter estimates and the corresponding standard errors for the first model are presented in the first part of Table 2.

The treatment seems to have a desired overall effect on both the level and the trend. We estimate that at the start of the second phase, the score decreases by more than one time the standard deviation of the scores around the regression lines. Moreover, the trend decreases by about one third of this standard deviation. On the basis of these overall parameters, one could predict, for a new case, the

evolution of the time series during the intervention phase, given his or her baseline scores. Using the Wald test and comparing the ratio of the parameter estimates and the corresponding standard errors with a standard normal distribution, we conclude that both $\hat{\gamma}_1$ and $\hat{\gamma}_3$ are significantly different from zero ($z = -4.64, p = .00005$, and $z = 27.021, p < .00001$, respectively).

The estimated values of σ_{u3}^2 and σ_{u13} are zero, which suggests that cases do not differ from each other in the effect on the trend. The estimate of the between-case vari-

ance of the coefficients indicating the effect on the level is 0.32. Yet, according to the Wald test with an alpha level of .05, this between-case variance is not significant ($z = 1.6, p = .11$).

A small between-case variance estimate of a certain effect suggests that the effect is similar for all cases. This means that the estimate of the overall effect can possibly be a better estimate of an individual effect than is the OLS estimate. Optimal estimates of the case-specific effects are obtained by using empirical Bayes techniques (Morris, 1983). These estimates are a weighted combination of the OLS estimates and the overall estimates. Although the empirical Bayes estimates are biased, using these estimates instead of the OLS estimates reduces the *MSE* (for a more elaborated discussion, see, e.g., Bryk & Raudenbush, 1992). In the example, the estimate of the variation of the treatment effects on the linear trend was zero. Therefore, the estimation of this effect is the same for each individual. This is not true for the effect on the level: Because the effect on the trend appeared to vary over cases, the estimated effects indeed vary around the overall effect, but are still “shrunk” to the overall effect.

There is a tendency for the effects on the level to vary from case to case. A possible explanation is that the cases did not receive the same treatment: Whereas some cases received behavioral counseling, others received client-centered counseling or no counseling. To explore the effect of the kind of treatment on the effect, we extend Equation 14 by regressing the effect on the level on dummies indicating the kind of treatment:

$$\hat{\beta}_{1j} = \beta_{1j} + e_{1j} \text{ and } \hat{\beta}_{3j} = \beta_{3j} + e_{3j},$$

and

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(\text{behavioral})_j + \gamma_{12}(\text{client})_j + u_{1j} \text{ and}$$

$$\beta_{3j} = \gamma_3 + u_{3j},$$

(15)

with $(\text{behavioral})_j$ and $(\text{client})_j$ equaling 1 if case j belongs to the behavioral counseling or the client-centered group, respectively, and 0 otherwise. While γ_{10} of Equation 15 represents the mean effect on the level in the no-contact

group, $(\gamma_{10} + \gamma_{11})$ and $(\gamma_{10} + \gamma_{12})$ represent the mean effect on the level in the behavioral and client-centered counseling groups, respectively.

The results are given in the second part of Table 2. The effect on the level in the behavioral counseling group appears to be more than twice the effect in the no-contact group (-1.52 vs. -0.70 ; $z = -2.65, p = .008$), whereas the effect in the client-centered group is almost twice the effect in the no-contact group, although the difference is not statistically significant at the .05 alpha level (-1.28 vs. -0.70 ; $z = 1.87, p = .06$). This conclusion is in line with the conclusion drawn by Marlowe et al. (1978) that “behavioral counseling, but not client-centered counseling was moderately helpful in reducing inappropriate classroom behavior” (p. 53). Our approach, however, gives a more detailed picture of the kind and the size of the effects. Finally, we note that the between-case variance of level effect disappeared. By taking the groups into account, the differences between the cases in the effect on the level thus are no longer larger than would be expected on the basis of the within-case variation alone.

Finally, we note that because, in the example, the raw data are all available, there is no need to work with effect size measures but that one could combine the raw data immediately, using hierarchical linear models (e.g., Verbeke & Molenberghs, 1997). A two-level model could be used, in which the coefficients (on the first level, the measurement level) of the regression equation of Center et al. (1985–1986) vary from case to case (on the second level). We found that the results are comparable with the results of using the effect size measure approach.

The Meta-Analysis for Other Single-Case Designs

In the preceding, we assumed that data stem from studies with an AB phase design. In practice, AB phase designs are of limited use, because they do not manage to control for possible influences of historical events or other sources of influence that change over time. As the following suggestions will show, the use of hierarchical linear models is a comprehensive approach that can be translated, modified, or applied to the specifics of more complex designs.

Table 2
Results of the Meta-Analyses of the Regression Coefficients

Parameter	Model 1		Model 2	
	Estimates	SE	Estimates	SE
Fixed Coefficients				
Effect on level				
Overall	γ_1	-1.16	0.25	—
No contact	γ_{10}	—	-0.70	0.26
Additional for behavioral counseling	γ_{11}	—	-0.82	0.31
Additional for client-centered counseling	γ_{12}	—	-0.58	0.31
Effect on trend	γ_3	-0.33	0.047	-0.33
			0.035	
(Co)variance Parameters				
Effect on level	σ_{u1}^2	0.32	0.20	0
Effect on trend	σ_{u3}^2	0	0	0
Covariance level slope	σ_{u13}	0	0	0

The models without the trend parameter can be used without modification for other phase designs with two treatments, as well as for alternation designs, as long as there are no linear trends and the performance is similar in all A phases and in all B phases. For ABAB designs with dissimilar performance in the A and/or B phases, we propose to use the following Level 1 equation for case j :

$$Y_{ij} = \beta_{0j}(\text{block1})_{ij} + \beta_{1j}(\text{phase})_{ij} * (\text{block1})_{ij} \\ + \beta_{2j}(\text{block2})_{ij} + \beta_{3j}(\text{phase})_{ij} * (\text{block2})_{ij} + e_{ij}, \quad (16)$$

with $(\text{phase})_{ij}$ being an indicator that equals 1 if occasion i is part of an intervention phase and 0 otherwise, and $(\text{block1})_{ij}$ and $(\text{block2})_{ij}$ being indicators that equal 1 if occasion i is part of the first or the second AB block, respectively, and 0 otherwise.

Whereas β_{0j} and β_{2j} indicate the mean performances in the A phases, β_{1j} and β_{3j} can be interpreted as the effects in the first and the second blocks, respectively. If the OLS estimates of the coefficients β_{1j} and β_{3j} can be obtained, a meta-analysis of these estimates can be performed, using a multivariate model with the effect size estimates for the different blocks as the dependent variables. To take possible linear trends into account, Equation 16 can be adapted in a similar way as that discussed above.

If phase or alternation designs are used to examine two or more treatments, one or more additional dummy variables can be included at the first level to indicate the treatment conditions. The results of a series of multiple baseline studies could be modeled by using a multivariate model, regarding the coefficients of the different response variables as different dependent variables, in case the same aspects are measured in the different studies. If the aspects that are investigated can rather be regarded as a random sample of possible aspects, an extra level can be included: Aspects are "nested" within cases.

Distributional Assumptions

In linear regression analysis, the estimation of the sampling distribution of the regression coefficients is based on the assumptions of independently normally distributed residuals and homogeneity of variance. The approximation of the sampling distribution of d is based on the same assumptions (Hedges, 1981). In addition, the approximation of the sampling distributions can be inadequate if based on a small number of observations. Moreover, although the sampling distribution of each regression coefficient divided by its standard error is a t distribution, in the present (multivariate) meta-analysis using hierarchical linear models, a (multivariate) normal sampling distribution of the effect sizes is assumed. Once more, this approximation is safe if the estimates are based on a sufficient number of observations.

According to Hedges and Olkin (1985), these assumptions are often justified in group comparison studies. However, in single-case studies, the number of data points is often small (Huitema, 1985). Moreover, because in a single-case experiment data are measurements stemming from the

same subject, serial dependence or autocorrelation may be present, violating the assumption of independence. Some claim that the existence of autocorrelation is a myth rather than a reality (e.g., Huitema, 1985), whereas others dispute this position (e.g., Matyas & Greenwood, 1991, 1997).

The influence of serial dependency on the research results and possible corrections for autocorrelation often receive little attention, especially beyond lag 1 autocorrelation (Matyas & Greenwood, 1997). However, most quantitative or qualitative analyses of single-case data are not immune to the possible existence of serial dependency (Gorman & Allison, 1997; Matyas & Greenwood, 1991, 1997). Some techniques for time series data have been developed to test for autocorrelation or to account for autocorrelation (e.g., Box & Jenkins, 1976; Gottman, 1981). If raw data are available, the regression models we presented for the calculation of effect sizes could be adapted in similar ways to account for possible autocorrelation in the data. Unfortunately, these procedures often raise problems, because many data points are necessary to evaluate and estimate autocorrelation in the data (Busse et al., 1995; Jayaratne, Tripodi, & Talsma, 1988). If all raw data are available, the use of hierarchical linear models offers an elegant solution to the problem of autocorrelation. Indeed, instead of assuming a simple covariance structure, a more complex covariance structure can be assumed that takes into account a possible autocorrelation, as is discussed in Verbeke and Molenberghs (1997). Whereas in regular approaches the autocorrelation is estimated for each case separately, when a hierarchical linear model is used a possible autocorrelation is estimated by using all data. In the example described above, we found that for none of the 12 time series was the Durbin-Watson statistic for a first-order autocorrelation significant at a .05 level. Similarly, the estimate of the first-order autocorrelation coefficient in the hierarchical linear model analysis of the raw data was zero.

Several alternatives are available for the meta-analysis of single-case data, with less restrictive assumptions about the sampling distributions of the effect size measures. One can, for instance, use the binomial test, the Wilcoxon model, or the one-sample t test to construct a confidence interval around the mean effect size (Busk & Serlin, 1992), use study sample sizes instead of the reciprocal of the sampling variance as weights in the fixed effects techniques of Hedges and Olkin (1985) for calculating the mean effect size (Faith, Allison, & Gorman, 1997), or combine p values, possibly resulting from permutation or randomization tests (Onghena, 1994; Onghena & Van Damme, 1994). Unfortunately, all of these alternatives have the disadvantage of not taking into account a possible linear trend. Moreover, the procedures do not allow answering all the important meta-analytic questions, including an estimate and explanation of the true effect size variance.

Conclusions

We showed that the results of single-case studies could be combined by performing a (multivariate) meta-analysis on (standardized) regression coefficients. In cases in

which there is no trend, this approach simplifies to a meta-analysis of standardized mean differences. Although the procedure outlined in this article does not solve all problematic aspects in the meta-analysis of single-case data, we tried to show that the procedure could give valuable information, which might be overlooked when other techniques are used.

The approach presented is indeed very appealing for four reasons. First, interventions often do not have constant effects, but effects are conditional on time. In contrast to traditional approaches, the use of a multivariate meta-analytic model in cases of linear trends allows us to maintain the specific information that is given by each coefficient and to determine differential effects of a moderator variable on the level and on the slope. Whereas a positive effect of a moderator variable on the mean performance can be obscured by an opposite effect on the trend if one effect measure is used, this is impossible if both kinds of effects are represented by different measures. Using two measures of effect size also means that predictions can be made about the effect of an intervention on a future occasion.

Second, the procedure easily extends to a meta-analysis of more than two regression coefficients—for instance, when the data for each case are modeled as a function of the time by means of a polynomial equation.

Third, hierarchical linear models are very flexible and can be adapted according to the specific meta-analytic data set and the research interests by including covariates on the different levels or by including additional levels. The models presented furthermore include the possibility that true effect sizes vary partly in a predictable way according to study characteristics, partly randomly. The general hierarchical linear meta-analytic model indeed includes fixed and random meta-analytic models, combining the strengths of both kinds of models (see National Research Council, 1992).

Finally, estimation procedures and user-friendly software are available to estimate and test the parameters of the models (for an overview, see Kreft & de Leeuw, 1998). In the estimation procedures, data are used efficiently in estimating unknown fixed and random parameters. Instead of the regression coefficients of the cases being estimated separately, empirical Bayes estimates of the individual coefficients use all data, resulting in more reliable estimates (Bryk & Raudenbush, 1992).

REFERENCES

- ALLISON, D. B., & GORMAN, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research & Therapy*, **31**, 621-631.
- BOX, G. E. P., & JENKINS, G. M. (1976). *Time series analysis, forecasting and control*. San Francisco: Holden-Day.
- BRYK, A. S., & RAUDENBUSH, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.
- BRYK, A. S., RAUDENBUSH, S. W., CONGDON, R. T., & SELTZER, M. (1988). *An introduction to HLM: User's guide, version 2.0* [Computer program]. Chicago: University of Chicago Press.
- BUSK, P. L., & SERLIN, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 187-212). Hillsdale, NJ: Erlbaum.
- BUSSE, R. T., KRATOCHWILL, T. R., & ELLIOTT, S. N. (1995). Meta-analysis for single-case consultation outcomes: Applications to research and practice. *Journal of School Psychology*, **33**, 269-285.
- CENTER, B. A., SKIBA, R. J., & CASEY, A. (1985-1986). A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education*, **19**, 387-400.
- COHEN, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.
- FAITH, M. S., ALLISON, D. B., & GORMAN, B. S. (1997). Meta-analysis of single-case research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 245-277). Mahwah, NJ: Erlbaum.
- GINGERICH, W. J. (1984). Meta-analysis of applied time-series data. *Journal of Applied Behavioral Science*, **20**, 71-79.
- GOLDSTEIN, H., RASBACH, J., PLEWIS, I., DRAPER, D., BROWNE, W., YANG, M., WOODHOUSE, G., & HEALY, M. (1998). *A user's guide to MLwiN*. London: University of London, Multilevel Models Project.
- 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, NJ: Erlbaum.
- GOTTMAN, J. M. (1981). *Time-series analysis: A comprehensive introduction for social scientists*. Cambridge: Cambridge University Press.
- HEDGES, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, **6**, 107-128.
- HEDGES, L. V., & OLKIN, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- HERSHBERGER, S. L., WALLACE, D. D., GREEN, S. B., & MARQUIS, J. G. (1999). Meta-analysis of single-case designs. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 107-132). London: Sage.
- HUITEMA, B. E. (1985). Autocorrelation in behavioral research: A myth. *Behavioral Assessment*, **1**, 107-118.
- HUITEMA, B. E., & MCKEAN, J. W. (2000). Design specification issues in time-series intervention models. *Educational & Psychological Measurement*, **60**, 35-58.
- JAYARATNE, S., TRIPODI, T., & TALSMAN, E. (1988). The comparative analysis and aggregation of single-case data. *Journal of Applied Behavioral Science*, **24**, 119-128.
- KALAIAN, H. A., & RAUDENBUSH, S. W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods*, **1**, 227-235.
- KREFT, I. G. G., & DE LEEUW, J. (1998). *Introducing multilevel modeling*. London: Sage.
- MARLOWE, R. H., MADSEN, C. H., JR., BOWEN, C. E., REARDON, R. C., & LOGUE, P. E. (1978). Severe classroom behavior problems: Teachers or counselors. *Journal of Applied Behavioral Analysis*, **11**, 53-66.
- MATYAS, T. A., & GREENWOOD, K. M. (1991). Problems in the estimation of autocorrelation in brief time series and some implications for behavioral data. *Behavioral Assessment*, **13**, 137-157.
- MATYAS, T. A., & GREENWOOD, K. M. (1997). Serial dependency in single-case time series. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 215-243). Mahwah, NJ: Erlbaum.
- MORRIS, C. (1983). Parametric empirical Bayes inference, theory and applications. *Journal of the American Statistical Association*, **78**, 47-65.
- NATIONAL RESEARCH COUNCIL (1992). *Combining information: Statistical issues and opportunities for research*. Washington, DC: National Academy Press.
- ONGHENA, P. (1994). *The power of randomization tests for single-case designs*. Unpublished doctoral dissertation, Katholieke Universiteit Leuven.
- ONGHENA, P., & VAN DAMME, G. (1994). SCRT 1.1: Single-case randomization tests. *Behavior Research Methods, Instruments, & Computers*, **26**, 369.
- PINHEIRO, J. C., & BATES, D. M. (2000). *Mixed-effects models in S and S-plus*. New York: Springer-Verlag.

- ROSENTHAL, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231-244). New York: Russell Sage Foundation.
- SHEU, C.-F., & SUZUKI, S. (2001). Meta-analysis using linear mixed models. *Behavior Research Methods, Instruments, & Computers*, **33**, 102-107.
- VERBEKE, G., & MOLENBERGHS, G. (1997). *Linear mixed models in practice: A SAS-oriented approach*. New York: Springer-Verlag.
- WANG, M. C., & BUSHMAN, B. J. (1999). *Integrating results through meta-analytic review using SAS software*. Cary, NC: SAS Institute.
- WHITE, D. M., RUSCH, F. R., KAZDIN, A. E., & HARTMANN, D. P. (1989). Applications of meta analysis in individual-subject research. *Behavioral Assessment*, **11**, 281-296.

(Manuscript received March 16, 2001;
revision accepted for publication March 19, 2002.)