British Journal of Mathematical and Statistical Psychology (1988), **41**, 257–278 *Printed in Great Britain*
© 1988 The British Psychological Society

Synthesizing standardized mean-change measures

Betsy Jane Becker
College of Education, Michigan State University, East Lansing, MI 48824-1034, USA

A new approach is presented for the meta-analysis of data from pre-test–post-test designs. With this approach, data from studies using different designs may be compared directly and studies without control groups need not be omitted. The approach is based on a 'standardized mean-change' measure, computed for each sample within a study, and involves analysis of the standardized mean changes and differences in the standardized mean changes. Analyses are illustrated using results of studies of the effectiveness of mental practice on motor-skill development.

1. Introduction

A new approach to the synthesis of data from pre-test–post-test designs is presented which allows results from several kinds of designs to be compared or combined directly, eliminating the need to omit studies because of design differences. Analyses are based on a 'standardized mean-change' measure, which is computed for each sample (experimental or control). Extensions of the analyses provide estimated experimental-versus-control comparisons for studies using the one-group pre-test–post-test design.

This new approach is valuable because it provides a way to accumulate information from more relevant studies than was previously possible. This should eliminate the arbitrary discarding or partitioning of meta-analysis data because of design differences and thereby promote more parsimonius syntheses of the existing literature on change.

Section 2 gives an overview of experimental designs utilizing pre- and post-testing. Then notation and a model for study results are introduced, the standardized mean-change measure is defined, and its sampling distribution is derived. Fixed-effects and mixed-model analyses for standardized mean changes from two-group studies are described in Section 4, and two approaches for including the results of no-control studies are given in Section 5. Section 6 outlines other problems which may be addressed through the use of standardized mean-change measures.

2. Research designs using pre-tests and post-tests

In many research areas one-group pre-test–post-test designs have been used to study the effects of experimental treatments or interventions. Change in average performance is attributed to the intervention. However, because several sources of invalidity are associated with this design (Campbell & Stanley, 1963) it is not widely applied.

The pre-test–post-test control group design compares the extent of change due to treatment to the change for an untreated (control) group (Campbell & Stanley, 1963). The control-group results provide an estimate of the amount of change due to retesting, historical artifacts, and growth (maturation). Other more elaborate designs involving pre- and post-testing (e.g. the Solomon four-group design) have also been used to study treatment effects .

Recently more sophisticated modelling techniques have been made available (e.g. Brogan & Kutner, 1980; Goldstein, 1979) which enable the analysis of treatment effects in longitudinal, multiwave data. Pre-test–post-test designs are the simplest (and least informative) form of longitudinal data-collection designs.

2.1. Synthesizing pre-test–post-test research

If several experimental designs are common to a single research literature a problem arises when the studies are reviewed quantitatively: Treatment effects in the designs are often represented by different statistical parameters and estimates. For example, the one-group design does not use a contrast between comparison groups to represent treatment effects, whereas most of the others do.

Glass's (1976) effect size, the measure of treatment effects used most often in quantitative research reviews, is the standardized difference between means of an experimental and a control group. Reviewers must either select from available research to use Glass's effect size (omitting studies without control groups), or use other research-synthesis methods such as vote counts or tests of combined results (e.g. Hedges & Olkin, 1985).

Omitting studies without control groups can involve considerable loss of data. For example, in a review of the literature on Scholastic Aptitude Test (SAT) coaching, DerSimonian & Laird (1983) listed seven samples without control groups, in a total of 22 samples. Studies without control groups are not always of poorer quality, but may simply address different questions (e.g. of differences in several treatments). For instance, many studies of the effects of mental practice on motor-skill learning (Feltz & Landers, 1983) compare mental-practice groups to control groups. However, some studies (e.g. Clark, 1960) compare skill improvement for mental-practice groups to that for groups experiencing actual physical practice of the skill. Such groups are obviously not comparable to control groups which do not practice, but are used to compare the effects of two types of practice, rather than the effect of one type of practice versus no practice at all.

Data from studies with and without control groups can be combined in meta-analysis by using standardized measures of mean change which are computed for each sample (experimental or control). The standardized mean change is the difference

between the post-test and pre-test means for one sample, divided by the pre-test (or post-test) standard deviation. (The consequences of using different standard deviations are discussed below). This measure is a scale-free index of the amount of change due to history, growth, and retesting for control groups; and of the effects of history, growth, retesting and treatment for experimental groups. Studies using pre-test–post-test control group designs thus produce two standardized mean changes, one for the experimental group and one for the control group.

The difference between the experimental- and control-group mean-change measures provides an estimate of the treatment effect. A modification of this difference in mean changes enables estimation of a treatment effect for studies using pre-test–post-test designs without control groups.

The methods described here were conceived as a solution to the problem of synthesizing existing studies on change. Many studies, especially older ones, have already collected pre-test–post-test data. More recent longitudinal studies and mult-wave panel data generally provide more information than pre-test–post-test data. Using only the first and last measures from such studies (for example) to compute standardized mean changes for a meta-analysis of multi-wave data could involve considerable loss of information. And the initiation of new studies using pre-test–post-test designs, simply for the purposes of synthesizing them using standardized mean changes, is definitely not to be recommended.

3. Notation and model for study results

Consider the results of the ith of k studies using pre-test (X) and post-test (Y) scores on a single outcome variable. In the simplest case each study has two samples, an experimental ($j = 1$) and a control ($j = 2$) group. Let X_{ijl} and Y_{ijl} be the pre-test and post-test scores, respectively, for the lth subject in group j in study i, and define n_{i1} and n_{i2} to be the numbers of experimental and control subjects, respectively, in study i. Assume that the pre-test and post-test scores are normally distributed with separate means and equal variances such that

$$\begin{aligned} X_{ijl} &\sim N(\mu_{ij}^X, \sigma_{ij}^2), \qquad \text{and} \\ Y_{ijl} &\sim N(\mu_{ij}^Y, \sigma_{ij}^2), \end{aligned} \tag{1}$$

for $l = 1, \ldots, n_{ik}; j = 1, 2$; and $i = 1, \ldots, k$. Here μ_{ij}^X and μ_{ij}^Y are the population means on X and Y for group j in study i, and σ_{ij}^2 is the population variance for both the pre-test and the post-test for group j in study i. Also define the pre-test–post-test correlation (between X_{ij} and Y_{ij}) to be ρ_{ij} for group j in study i.

The assumption that the pre-test and post-test variances are equal within the jth sample of study i simplifies the sampling distribution of the standardized mean change as it is defined below. Glass, McGaw & Smith (1981), and McGaw & Glass (1980) made this same assumption in their discussions of the computation of effect sizes based on gain scores for experimental and control groups.

3.1. Definition of the standardized mean change

The standardized mean change is the standardized mean difference between post-test and pre-test averages for a single sample. The sample *standardized mean change* is defined as

$$g_{ij} = \frac{\bar{Y}_{ij} - \bar{X}_{ij}}{S_{ij}}, \tag{2}$$

where $\bar{X}_{ij}$ and $\bar{Y}_{ij}$ are the sample pre-test and post-test means and S_{ij} is the pre-test (or post-test) standard deviation for sample j in study i. The standardized mean-change measure represents average change from pre-test to post-test in standard-deviation units. The standardized mean-change measure is not identical to Glass's effect size, which has also been denoted as g. The symbol g (rather than g^C, for example) is used here for simplicity though the estimators are different.

Under the model for scores proposed in (1), the variances of X and Y are equal. Thus if the model is true, the standard deviation based on either X or Y may be used as S_{ij}. If the treatment or the circumstance of repeated testing is thought to influence the variability of scores, the pre-test standard deviation provides a more sensible value for S_{ij}, since this measure would presumably be free of either influence.

Two alternative statistics that could be used to synthesize data from pre-test–post-test studies are the mean gain and mean residual gain for each group divided by the standard deviation of the gain scores. When studies report gain or residual scores (rather than results for both pre- and post-tests) these may be the only calculable measures, and residual gains in particular do not suffer from some of the problems of simple change scores. However, such indices are more difficult to interpret than the standardized mean change shown above because they are not in the scale of the original scores. Also residual gain analyses have been used less frequently, and standardized mean residual gains often can not be determined from the data available in published study reports.

The sample standardized mean change estimates a population parameter, denoted δ_{ij}, where

$$\delta_{ij} = \frac{\mu_{ij}^{Y} - \mu_{ij}^{X}}{\sigma_{ij}}, \tag{3}$$

and μ_{ij}^{Y}, μ_{ij}^{X}, and σ_{ij} are defined as for (1) above.

One limitation of the standardized mean change is its reliance on measures at only two time points. Although often used in educational research, two-wave designs (and consequently standardized mean changes) do not allow one to distinguish linear from non-linear trends. A second, related weakness is the fact that the standardized mean-change measure does not account for the amount of time between pre- and post-testing. Two samples may exhibit the same amount of change yet if the time between testing sessions differs for the samples, the interpretation of the two equal-sized mean-change measures may differ.

3.2. Distribution of the standardized mean change under fixed effects

The sample standardized mean change g is similar to Glass's original effect size comparing means for treatment and control groups (i.e. it is a mean difference divided

by a standard deviation), yet because it is based on paired or repeated-measure data from a single sample its sampling distribution is different.

The results presented first are based on the assumption that the population standardized mean changes can be characterized by a fixed-effects model. That is, the parameters are presumed to be fixed, thus all variability in sample mean changes is considered to result from sampling error rather than variation in the true values of the δ_{ij}. Alternatives to the fixed-effects model are discussed below.

Result 1. Suppose that $X_{ijl} \sim N(\mu_i^X, \sigma_{ij}^2)$ and $Y_{ijl} \sim N(\mu_{ij}^Y, \sigma_{ij}^2)$ as in model (1), with $\delta_{ij} = (\mu_{ij}^Y - \mu_{ij}^X)/\sigma_{ij}$ and ρ_{ij} as the pre-test–post-test correlation. Then g_{ij} for $j = 1, 2$; $i = 1, \ldots, k$ has exact mean and variance given by

$$E(g_{ij}) = \frac{\delta_{ij}}{c(n_{ij}-1)}, \tag{4}$$

and

$$\text{var}(g_{ij}) = \frac{2(1-\rho_{ij})}{n_{ij}}\left[\frac{(n_{ij}-1)}{(n_{ij}-3)}\left\{1+\frac{n_{ij}(\delta_{ij})^2}{2(n_{ij}-1)}\right\} - \frac{(\delta_{ij})^2}{c(n_{ij}-1)}\right], \tag{5}$$

where $c(m)$ is defined by

$$c(m) = \sqrt{\frac{2}{(m-1)}}\frac{\Gamma[(m)/2]}{\Gamma[(m-1)/2]},$$

and approximated (Hedges, 1981) by $c(m) \approx 1 - 3/(4m-1)$. An approximate formula for the variance of g_{ij} is given by

$$\begin{aligned}\text{var}(g_{ij}) &= \frac{2(1-\rho_{ij})}{n_{ij}}\left\{1+\frac{(n_{ij}(\delta_{ij})^2/2(1-\rho_{ij}))}{2(n_{ij}-1)}\right\} \\ &= \frac{2(1-\rho_{ij})}{n_{ij}} + \frac{(\delta_{ij})^2}{2(n_{ij}-1)}.\end{aligned} \tag{6}$$

Proof. The theoretical variance of the mean difference $(\bar{Y}_{ij} - \bar{X}_{ij})$ is $2(1-\rho_{ij})\sigma_{ij}^2/n_{ij}$, where ρ_{ij} is the population pre-test–post-test correlation and where both the pre-test and post-test variances equal σ_{ij}^2 in group j of study i. Thus the standardized mean-change measure g_{ij} may be written as a constant times the ratio of a unit normal variable with a non-zero mean (that is, the ratio of $\sqrt{n_{ij}}(\bar{Y}_{ij} - \bar{X}_{ij})/\sqrt{2(1-\rho_{ij})}\sigma_{ij}$) to the square root of a chi-square variable with $(n_{ij}-1)$ degrees of freedom (i.e. $\sqrt{S^2/\sigma^2}$). Specifically,

$$\sqrt{\frac{n_{ij}}{2(1-\rho_{ij})\sigma_{ij}}}\cdot\left(\frac{\bar{Y}_{ij} - \bar{X}_{ij}}{S_{ij}/\sigma_{ij}}\right) = \sqrt{\frac{n_{ij}}{2(1-\rho_{ij})}}g_{ij}. \tag{7}$$

Since (7) is a non-central t variate with $(n_{ij}-1)$ degrees of freedom and non-centrality parameter $\sqrt{n_{ij}/2(1-\rho_{ij})}\delta_{ij}$, the standardized mean change g_{ij} is distributed as $\sqrt{2(1-\rho_{ij})/n_{ij}}$ times a non-central t variate with non-centrality parameter $\sqrt{n_{ij}/(2(1-\rho_{ij}))}\delta_{ij}$.

The moments for g_{ij} are then obtained from the exact and approximate moments for the non-central t distribution, provided, for example, by Johnson & Kotz (1970, pp. 203–204). ■

Result 2. If the g_{ij} are the estimators defined in (2), then as $n_{ij} \to \infty$, for $j = 1, 2$; $i = 1, \ldots, k$, the distribution of the standardized mean-change measure tends to a normal distribution such that

$$\sqrt{n_{ij}}(g_{ij} - \delta_{ij}) \sim N(0, [2(1 - \rho_{ij}) + (\delta_{ij})^2/2]). \tag{8}$$

Proof. The distribution is obtained as a consequence of the large-sample normal approximation to the non-central t distribution. ■

Observations concerning g_{ij}. Because the expected value of g_{ij} is not equal to δ_{ij}, g_{ij} is a biased estimator of the population standardized mean change. However, as $n_{ij} \to \infty$, the value of the function $c(n_{ij} - 1)$ approaches 1.0, thus g_{ij} is asymptotically unbiased.

The standardized mean-change measure defined in (2) is a consistent estimator of δ_{ij}. First, as $n_{ij} \to \infty$ the expectation of g_{ij} approaches δ_{ij}. Also as n_{ij} increases, the variance of the standardized mean change tends to zero. [This is most obvious in the approximate formula for the variance given in (6).] Thus the estimator is consistent.

Though the estimator g_{ij} is consistent, it is not asymptotically efficient. This is seen by comparing the variance of g_{ij} to the theoretical minimum variance. The theoretical minimum variance (Cramer–Rao bound) is given by the variance of the maximum-likelihood estimator of δ, denoted here as $\hat{\delta}$. The bivariate likelihood of X and Y was used to obtain this variance, assuming that all parameters were to be estimated (that is, no parameters were assumed to be known). The Cramer–Rao bound is

$$\text{var}(\hat{\delta}_{ij}) = \frac{2(1 - \rho_{ij})}{n_{ij}} + \frac{(\delta_{ij})^2(1 + \rho_{ij}^2)}{4n_{ij}}.$$

The ratio of the variance of the standardized mean-change measure to the Cramer–Rao bound is

$$\frac{\text{var}(g_{ij})}{\text{var}(\hat{\delta}_{ij})} = \frac{2(1 - \rho_{ij}) + (\delta_{ij})^2/2}{2(1 - \rho_{ij}) + (\delta_{ij})^2(1 + \rho_{ij}^2)/4}.$$

Because the ratio of the variances does not approach one as $n_{ij} \to \infty$, the estimator g_{ij} is not asymptotically efficient in general. However, when the value of the pre-test–post-test correlation is either 1.0 or −1.0, the ratio equals one. Thus when there is a perfect correlation between X and Y, the estimator g_{ij} is asymptotically efficient.

More generally, when there is a fairly strong relationship between the pre-test and post-test, the variance of the standardized mean-change measure is near the minimum (that is, g_{ij} is almost fully efficient). Precision is at a minimum when the pre-test and post-test are unrelated in the population (i.e. when ρ is 0).

When the variances of the X and Y *scores* differ, the exact variance of the standardized mean-change measure computed using the pre-test standard deviation is more complicated than that shown above. However, if the post-test variance is not dramatically larger or smaller than the pre-test variance, the variance of the

standardized mean change given above will be a reasonable approximation to the proper value.

3.3. An unbiased standardized mean-change measure

Since the expectation of g_{ij} is a constant times the parameter δ_{ij}, g_{ij} is a biased estimator. However, an unbiased estimate denoted d_{ij} can be obtained by multiplication. Specifically,

$$d_{ij} = c(n_{ij} - 1)g_{ij}, \tag{9}$$

for $j = 1, 2$; $i = 1, \ldots, k$, and where $c(n_{ij} - 1)$ is given for (4) and (5) above.

The exact variance of d_{ij} may be obtained from the variance of g_{ij} given in (5) above. $\mathrm{Var}(d_{ij})$ is given by

$$\mathrm{var}(d_{ij}) = [c(n_{ij} - 1)]^2 \,\mathrm{var}(g_{ij}), \tag{10}$$

thus $\mathrm{var}(d_{ij})$ is always less than $\mathrm{var}(g_{ij})$. It is usually useful to approximate $\mathrm{var}(d_{ij})$ by

$$\mathrm{var}(d_{ij}) = \frac{2(1 - \rho_{ij})}{n_{ij}} + \frac{(\delta_{ij})^2}{2n_{ij}}. \tag{11}$$

Because d_{ij} is an unbiased estimator of the population standardized mean change, and because its variance is less than that for g_{ij}, the estimator d_{ij} is henceforth used as the standardized mean-change measure.

Result 3. If d_{ij} is the unbiased estimator defined in (9) then as $n_{ij} \to \infty$, the distribution of d_{ij} tends to a normal distribution such that

$$\sqrt{n}_{ij}(d_{ij} - \delta_{ij}) \sim N(0, [2(1 - \rho_{ij}) + (\delta_{ij})^2/2]) \tag{12}$$

for $j = 1, 2$; $i = 1, \ldots, k$.

Proof. As $n_{ij} \to \infty$, the distribution of d_{ij} tends to that of g_{ij}, which is the distribution shown in (8) and (12). ■

3.4. Estimating the variance of the standardized mean-change measure

The variance of the standardized mean change is a function of the population mean change as well as of the population pre-test–post-test correlation. In general, however, values of δ and ρ are not known. The variance thus must be computed using estimates of these parameters. If consistent estimates of δ and ρ are available, then as $n \to \infty$, the asymptotic variance (computed by using estimates of δ and ρ) will equal the variance stated in terms of parameters.

The computational formula used to estimate the variance of d_{ij} is

$$\mathrm{var}(d_{ij}) = \frac{2(1 - r_{ij})}{n_{ij}} + \frac{(d_{ij})^2}{2n_{ij}}, \tag{13}$$

where d_{ij} is defined in formula (9) and r_{ij} is the best available estimate of the pre-test–

post-test correlation for group j of study i. Even when the values of pre-test–post-test correlations are not presented explicitly one may still be able to estimate them from t statistics computed using change scores or an average estimate may be obtainable from the analysis-of-variance F test.

4. Analyses of standardized mean changes

Because the asymptotic distribution of the standardized mean change is very similar to that of Glass's effect size, mean changes from separate samples could be considered independent data points and analysed using methods of meta-analysis devised for effect-size data (e.g. Hedges & Olkin, 1985; Rosenthal & Rubin, 1982). Standardized change measures would be accumulated across studies within treatment and control populations, and analyses of the independent treatment and control effects could explore relationships of relevant predictors to the standardized mean changes.

One might argue for such an approach because, on the basis of the usual assumptions (e.g. for the t test), experimental and control samples within each study are assumed to be statistically independent (unless matching has been used). However, this argument overlooks one implicit assumption of the usual analyses: the groups being compared within each study are assumed to be similar (on all important variables) except that one has received a treatment of some kind. The reason that the within-study comparisons are useful is that the groups *within studies* are similar.

Thus it is not sensible to ignore these implicit similarities among the groups from each study. An approach is needed that relates and compares to each other the outcomes within each study. Several analyses can be designed to that end; the fixed-effects approaches described here focus on the analysis of within-study differences in mean-change measures, and the mixed-model analyses examine separate group effects at the within-study level and differences between experimental and control parameters at the between-study level.

Fixed-effects analyses of within-study differences are first described, for the simple situation in which all studies are of two-group comparisons. Second an approach based on a mixed model for the experimental-control pairs of standardized mean changes is explored.

4.1. Differences in standardized mean changes for two-group comparison studies

Consider the case in which each study to be synthesized compares a single experimental group to a single control group. Each study has two standardized mean changes. The parameter representing the difference in average change between the treatment and control groups in the ith study is

$$\Delta_i = \delta_{i1} - \delta_{i2}, \tag{14}$$

where δ_{i1} and δ_{i2} represent the population standardized mean-change measures for the ith experimental ($j = 1$) and control ($j = 2$) groups, respectively. The parameter Δ thus symbolizes change due to treatment with the effects of history, retesting, and maturation removed.

The 'scale' of Δ is standard-deviation units, as is true for the standardized mean changes. However, the value of Δ represents the difference between the number of standard deviations of change in the experimental and control groups. It is possible for neither group to exhibit the degree of change expressed by Δ because Δ is a difference. Only if the ith control population shows no change (i.e. $\delta_{i2} = 0$) is the ith experimental standardized mean change equal to Δ_i.

The difference $\hat{\Delta}_i$ between the unbiased sample mean changes is an unbiased estimator of the parameter Δ_i. The sample difference is

$$\hat{\Delta}_i = d_{i1} - d_{i2}. \tag{15}$$

The variance of the difference $\hat{\Delta}_i$ is simply the sum of the variances for the two standardized mean-change estimates because the two mean changes are computed from independent samples within study i. Thus the variance of the difference is

$$\text{var}(\hat{\Delta}_i) = \text{var}(d_{i1}) + \text{var}(d_{i2}), \tag{16}$$

where $\text{var}(d_{ij})$ is given by (13).

The result shown in (12) implies that the distribution of $\hat{\Delta}_i$ is also approximately normal. For the ith study, define $n_i = n_{i1} + n_{i2}$, and let $\pi_{ij} = n_{ij}/n_i$ be fixed as $n_i \to \infty$ for $j = 1, 2$; and $i = 1, \ldots, k$. Then as n_i approaches ∞,

$$\sqrt{n_i}(\hat{\Delta}_i - \Delta_i) \sim N(0, \text{var}(d_{i1})/\pi_{i1} + \text{var}(d_{i2})/\pi_{i2}), \tag{17}$$

where $\text{var}(d_{ij})$ is estimated by (13).

The distribution of $\hat{\Delta}_i$ has the same form as that described by Hedges (1983*b*), thus the analyses described by Hedges for such estimators can be directly applied to the differences in standardized mean-change measures. To avoid duplicating the presentation of test statistics and estimates in Hedges (1983*b*), only Hedges' formula for the test of homogeneity is given. The analyses are applied here in examples using estimates of change in motor-skill performance from studies of mental practice.

4.2. Example of two-group comparisons

The studies. The data are drawn from a meta-analysis of studies on the effectiveness of mental practice on motor-skill development (Feltz, Landers & Becker, 1986). Table 1 lists sample sizes, unbiased standardized mean changes and their estimated variances, and a description of the motor task measured in each study.

The first five studies provided data on two-group comparisons. The last three used either single-group pre-test–post-test designs or multi-group designs with comparison groups which experienced some kind of practice (i.e. not *control* comparison groups). .The results of these three studies are used in further analyses below.

Consistency of the differences. The first hypothesis tested was the hypothesis of consistency or homogeneity of the differences (the $\Delta_i$values). The null hypothesis was

$$H_0: \Delta_1 = \Delta_2 = \cdots = \Delta_k = \Delta.$$

This model suggests that the advantage of mental practice versus no practice for changing motor-skill performance is the same in all studies. Tests of the model do not

Table 1. Results of studies of mental practice and motor-skill performance

	Mental practice			Control			
Study	*n*	*d*	var(*d*)	*n*	*d*	var(*d*)	Measure type
Kovar (1969)	16	0.55	0.055	16	0.85	0.063	Underhand frisbee
Perry (1939)	16	0.84	0.068	16	0.65	0.053	Three-hole tracing
Rawlings & Rawlings (1974)	18	2.29	0.195	12	1.86	0.209	Pursuit rotor
Ryan & Simons (1982)	8	0.49	0.107	16	−0.05	0.039	Stabilometer
Whitehill (1964)	19	0.83	0.057	19	0.05	0.033	Handball serve
Bissonette (1965)	10	0.62	0.094	—	—	—	Speed skating
Standridge (1971)	10	0.25	0.075	—	—	—	Swimming whip kick
Tufts (1963)	13	0.30	0.059	—	—	—	Bowling
Weighted average		0.65			0.36		
Standard error		0.10			0.10		

Note. Sample sizes (*n*), standardized mean changes (*d*), and variances (var(*d*)).

indicate whether there is an advantage for mental practice, but only examine agreement among the results.

Hedges (1983*a*, p. 125) provided a test of the homogeneity of a set of parameters which can reasonably be assumed to be fixed, in this case, the Δ_i values. The statistic has an asymptotic distribution which is chi-square with $k-1$ degrees of freedom when the null hypothesis is true. Using the present notation, Hedges' statistic is

$$\sum_{i=1}^{k} \frac{\hat{\Delta}_i^2}{\text{var}(\hat{\Delta}_i)} - \left[\sum_{i=1}^{k} \frac{\hat{\Delta}_i}{\text{var}(\hat{\Delta}_i)}\right]^2 \Big/ \sum_{i=1}^{k} \frac{1}{\text{var}(\hat{\Delta}_i)}.$$

The first two columns of Table 2 present the standardized mean-change measures and their variances for the five studies which examined both a mental-practice and a control group. Because the model for scores on X and Y allowed the correlation to differ both between and within studies, one approach for computing the variances would have been to substitute sample correlations (i.e. the r_{ij} values) for the ρ_{ij} values. However, because few studies presented data sufficient for estimating the pre-test–post-test correlations, common estimates of the correlations (based on all available *r*s reported in the review) were used. The correlation used in the mental-practice variances was 0.64 and that used in the control-group variances was 0.69.

For the five two-group studies, Hedges' homogeneity-test value was 6.19. Compared to percentage points of the chi-square distribution with $(k-1) = 4$ degrees of freedom, the test was not significant at the $\alpha = 0.05$ level, suggesting the differences in standardized mean changes were reasonably similar across studies.

Magnitude of the difference in change. The next question concerned the magnitude of the common (or average) value of Δ for the studies. (When study results are not

Table 2. Differences in standardized mean-change measures

Study	Two-group studies		All studies[a]	
	$\hat{\Delta}_i$	var($\hat{\Delta}_i$)	$\hat{\Delta}_i$	var($\hat{\Delta}_i$)
Kovar (1969)	−0.30	0.12	−0.30	0.12
Perry (1939)	0.19	0.12	0.19	0.12
Rawlings & Rawlings (1974)	0.43	0.40	0.43	0.40
Ryan & Simons (1982)	0.54	0.15	0.54	0.15
Whitehill (1964)	0.79	0.09	0.79	0.09
Bissonette (1965)	—	—	0.62	0.09
Standridge (1971)	—	—	0.25	0.08
Tufts (1963)	—	—	0.30	0.06
Weighted average	0.34		0.35	
Standard error	0.16		0.11	

[a] The estimates of Δ_i for the last three studies were computed by substituting zero for missing control-group results.

homogeneous there is no *common* parameter, though an average may be computed.) An estimate of Δ, given by the weighted average Δ_i estimate, was 0.34 with a standard error of 0.16. A 95 per cent confidence interval for Δ was from 0.02 to 0.66, thus the control and mental-practice groups differed significantly in degree of change in motor skill. The positive interval indicates that each mental-practice group can expect to gain in skill *more* than a group experiencing no practice.

The differences in standardized mean changes in motor skill agreed with the model of a single effect for the mental-practice treatment gain. An increase of about one-third of a standard deviation *beyond* what would be expected from simply repeating the task was gained from the use of mental practice. On a motor-skill measure with a standard deviation of 10 points, an average subject using mental practice would be expected to gain almost three and one-half points more than an average control subject.

4.3. Further analyses of the differences in standardized mean changes

When homogeneity tests indicate inconsistency of the fixed-effects parameters or when the reviewer has *a priori* hypotheses about relationships of study features to the sizes of experimental–control differences in change, further analyses are needed. The analyses utilize the Δ_i estimates and proceed as outlined by Hedges (1983*b*).

The relationships to be analysed in a particular review depend on the phenomenon under study. However, two predictors should routinely be considered if information about them is available. The *time interval* between pre-testing and post-testing is an important explanatory variable for variation in amounts of change, especially when the outcome represents physical growth or another construct for which change due to natural maturation is expected. Individuals studied for different durations may be expected to show different amounts of growth, independent of the influences of

treatments. Similarly for some outcomes the *amount of practice* between pre-test and post-test (independent of the time interval between testings) may be relevant.

4.4. Alternative fixed-effects analyses

There are several alternatives to the analyses described above. One which was mentioned above is to treat each standardized mean change as an independent data point and ignore within-study associations. However, this analysis is undesirable because it ignores implicit within-study similarities.

A second alternative would be to examine the unweighted mean difference between the experimental and control change measures. This analysis has the advantage of always providing an unbiased estimate of $\delta_{i1} - \delta_{i2}$. In the case of inconsistent Δ values, the unweighted mean difference is the preferred estimate of $\delta_{i1} - \delta_{i2}$ because it does not give more weight (i.e. more influence) to larger samples. However, when the Δ_is are consistent, the unweighted estimate is less precise.

4.5. Mixed-model analyses

More complex methods of analysing the standardized mean changes might use random-effects (e.g. Hedges, 1983*a*) or mixed-model (Raudenbush & Bryk, 1985) approaches. The population standardized mean changes could be assumed to vary both within and between studies, and relationships between experimental and control population results could be modelled as well. A hierarchical mixed-model approach is used in the following example.

Model. A simple model would begin with the same assumptions made above about the distributions of effects within studies. This is equivalent to assuming that each standardized mean change d_{ij} is composed of a population parameter δ_{ij} and some component of error e_{ij}. That is, $d_{ij} = \delta_{ij} + e_{ij}$ and each e_{ij} is assumed to be normally distributed with a mean of 0 and a variance var(d_{ij}) given in (10).

The mixed-model analysis diverges from the fixed-effects approach by assuming that each δ_{ij} is also a random variable, composed of a component that is fixed across studies (say γ_j) and an error component u_{ij}. The parameter γ_1 would represent the average or common experimental-group or treatment effect across studies, and γ_2 would be the average control-group standardized mean change. Thus

$$\delta_{ij} = \gamma_j + u_{ij}, \qquad \text{for } j = 1, 2, i = 1, \ldots, k,$$

and the vectors of errors $\mathbf{u}_i = (u_{i1}\ u_{i2})'$ for $i = 1, \ldots, k$, are assumed to be normally distributed with a mean vector of $\mathbf{0}$ and with common (across studies) variance–covariance matrix $\mathbf{T}$. Formally, for $i = 1, \ldots, k$,

$$\begin{bmatrix} u_{i1} \\ u_{i2} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}\right).$$

Because each δ_{ij} is composed of a fixed part plus the error u_{ij}, the $\mathbf{T}$ matrix also is the variance–covariance matrix for the vector of effects $\boldsymbol{\delta}_i = (\delta_{i1}\ \delta_{i2})'$.

Combining the two models allows d_{ij} to be written as

$$d_{ij} = \gamma_j + u_{ij} + e_{ij},$$

which shows that each d_{ij} consists of a fixed part γ_j and two sources of error, within-study sampling error represented by e_{ij}, and parameter variability, represented by u_{ij}. The random nature of δ_{i1} and δ_{i2} influences the standardized mean-change estimates via the u_{ij} term. Also, this error term introduces both parameter variability and covariation (across studies) between the experimental and control δ values into the estimates of variability for each standardized mean-change estimate d_{ij}.

Example. On the basis of the above model the standardized mean changes for the five studies having results for both experimental and control groups were analysed using the Hierarchical Linear Model (HLM) program (Bryk, Raudenbush, Seltzer & Congdon, 1986). The HLM program uses the EM algorithm and empirical Bayes procedures for the estimation of parameters. Bryk *et al.* (1986) describe the details of the estimation procedures and other applications of the HLM approach.

The average experimental-group and control-group standardized mean changes (the γ parameters) were estimated from the d_{ij} values, weighted in proportion to their precision matrix. The average change for experimental groups was estimated to be $\hat{\gamma}_1 = 0.94$ standard-deviation units (with standard error SE = 0.30), and for control groups it was $\hat{\gamma}_2 = 0.63$ (SE = 0.34).

The estimate of the variance–covariance matrix $\mathbf{T}$ was

$$\hat{\mathbf{T}} = \begin{bmatrix} 0.351 & 0.387 \\ 0.387 & 0.518 \end{bmatrix}.$$

Both parameter variances (the diagonal elements of $\mathbf{T}$) differed significantly from zero, indicating considerable inconsistency among the δ_{ij} values. (The chi-square statistic for the test that $T_{11} = 0$ was 14.40, d.f. = 4, $p = 0.006$; for the test that $T_{22} = 0$ the chi-square was 30.25, d.f. = 4, $p = 0.0005$).

The correlation between the experimental and control standardized mean changes across studies (computed from the $\hat{\mathbf{T}}$ elements) was 0.91, which is quite high. Within-study similarities in treatment implementation, selection of outcome measures, and choice of subject population have apparently created a strong relationship between δ_{i1} and δ_{i2} across studies. In analyses of larger data sets some of this intercorrelation may be accounted for through either fixed-effects or mixed-model analyses that examine relationships of between-study differences to study outcomes. However, only the mixed-model analysis provides an approach to assessing the degree of remaining interrelatedness.

The product moment correlation for the pairs of d_{ij} values, an estimate that includes not only covariation between the δ_i values but also sampling variability, might be used as an approximate indicator of parameter intercorrelation when standardized mean changes arises from large samples. The unweighted correlation between the experimental and control d values for the example data was high, with $r = 0.85$.

A primary goal of the analysis was to determine whether change for the experimental groups was greater than change for the control groups. This can be

examined via a test of the hypothesis that $\gamma_1 = \gamma_2$ (or $\gamma_1 - \gamma_2 = 0$). In this test the elements of the variance–covariance matrix of $\hat{\gamma}$ are used to estimate the variance of the difference $\hat{\gamma}_1 - \hat{\gamma}_2$, which is $\text{var}(\hat{\gamma}_1) + \text{var}(\hat{\gamma}_2) - 2\,\text{cov}(\hat{\gamma}_1, \hat{\gamma}_2)$. The test, which is a standard normal variate under H_0, is

$$z = \frac{\hat{\gamma}_1 - \hat{\gamma}_2}{\sqrt{\text{var}(\hat{\gamma}_1 - \hat{\gamma}_2)}}. \tag{18}$$

The variance–covariance matrix of the γ_j estimates, $\mathbf{V}(\hat{\gamma})$ is obtained from $\mathbf{T}$ and the variance matrices for the individual studies. Specifically,

$$\mathbf{V}(\hat{\gamma}) = \left[\sum_{i=1}^{k} \begin{bmatrix} T_{11} + \text{var}(d_{i1}) & T_{12} \\ T_{21} & T_{22} + \text{var}(d_{i2}) \end{bmatrix}^{-1}\right]^{-1}. \tag{19}$$

The estimate based on the example data was

$$\hat{\mathbf{V}}(\hat{\gamma}) = \begin{bmatrix} 0.087 & 0.079 \\ 0.079 & 0.117 \end{bmatrix},$$

so that the standard error of the difference between the experimental and control γ estimates was 0.218. The value of the test statistic (18) for the example data was $z = 1.42$, which was not significant at even the lenient 0.10 level. The hypothesis of equal standardized mean changes for experimental and control populations was accepted.

The hierarchical analysis suggested that on average the subjects experiencing mental practice gained 0.31 standard-deviation units more on measures of motor-skill than did control subjects. This is very close to the estimate of 0.34 obtained above from the $\hat{\Delta}_i$ values. In the mixed model analysis, however, the advantage was not statistically significant. The standard error in the denominator of the test (18) includes both sampling variation and parameter variability (as indicated in (19)), and analysis of the elements of $\mathbf{T}$ identified significant parameter variability in the δ_{ij}s. This variability increased the standard error in (18), and was not entirely counterbalanced by the high intercorrelation between δ_{i1} and δ_{i2}.

4.6. Fixed-effects versus random-effects or mixed models

The decision of whether to apply a fixed-effects, random-effects, or mixed model in an analysis of standardized mean-change measures is made in the same way as decisions concerning the application of such models in analysis of variance. If the researcher believes that there is likely to be a distribution of 'true' effects for experimental and for control groups, rather than shared effects for all treatment and control populations, then random or mixed models are more appropriate. If the researcher thinks that the true values of shared population parameters can be determined through the knowledge of only one or a few study characteristics, then fixed-effects analyses may be appropriate.

If there is no variation in the distributions of parameters hypothesized to be random, the random-effects model simplifies to the fixed-effects model. Thus the random and mixed models are more general. However, they are somewhat more

difficult computationally, whereas the fixed-effects analyses can be accomplished via common statistical packages such as SAS and SPSS.

5. Fixed-effects analyses including studies without control groups

Two approaches can be imagined for including the results of no-control studies into analyses of standardized mean changes. One approach is to substitute some 'reasonable' constant value for the control standardized mean-change measure, and the other is to estimate such a value. The two approaches have different consequences for the variances of the $\hat{\Delta}$s and their analysis.

Regardless of the approach used to include studies without control groups, it is useful to compare the results of such no-control studies to those of the two-group studies. Differences in results or an absence of differences may result from the use of inappropriate substitute values for the control-group effects, or may arise because of design inadequacies or other substantively interesting differences between the two kinds of study.

Whether the no-control and two-group studies differ cannot be adequately determined by any statistical test. Further studies may be needed to determine the nature of empirically identified differences. However, if the two kinds of study appear similar on all important characteristics (other than the presence of a control group), and if their results appear similar when reasonable values are substituted for the non-existent control-group results, the reviewer has no empirical grounds for discarding or separating the results from the no-control studies in the review at hand.

Because of the implicit question of differences in quality between the two-group and no-control studies, the reviewer should carefully consider all study features and identify variables which are confounded with absence of a control group (especially variables which are substantively more interesting than the lack of a control group). Identifying such variables reduces the chance of making unjustified claims about substantive study features that are actually confounded with design features, and also may identify questions for future investigation.

5.1. Using a constant for the control-group value of δ

Using a constant for the control group's standardized mean-change measure is the simpler approach. The reviewer selects, on the basis of a rational argument or previous information, a single value C to represent the expected amount of change due to retesting and maturation. The value C is then used to compute $\hat{\Delta}_i$ as $d_{i1} - C$ for studies without control groups.

Imagine that no improvement due to retesting or maturation is expected for motor skills. Then zero would be a reasonable value for the expected change (C) for a control group. In a review of the effectiveness of SAT coaching one might use as the 'missing' standardized mean change the number of points typically gained on retesting, expressed in standard-deviation units. If 50 points were the typical expected gain from retesting on the SAT, using the population standard deviation of 100 points, $C = 0.5$ would be used for the control-group standardized mean-change measure.

Because the value used for the missing control-group results is a constant, it does not influence the variances of the estimates of Δ in the no-control studies. The variance of $\hat{\Delta}_i$ for a study with a constant C used for its non-existent control-group mean change is

$$\text{var}(\hat{\Delta}_i) = \text{var}(d_{i1}) + \text{var}(C) = \text{var}(d_{i1}). \tag{20}$$

One problem with this result, though it is technically correct, is that if the result of exactly zero had been obtained by a sample of control subjects, it would be associated with some sampling variability. This sampling variability would increase the value of $\text{var}(\hat{\Delta}_i)$ in proportion to the number of subjects in the control group. A somewhat more conservative (though *ad hoc*) estimate of the sampling variance of the imputed value could be obtained by assuming that if a control group had been used it would have been the same size as the experimental groups from the study. Then the *ad hoc* variance of the substituted value (denoted $\bar{\delta}_{i2}$) could be computed as

$$\text{var}(\bar{\delta}_{i2}) = \frac{2(1-\bar{\rho}_{i2})}{n_{i1}} + \frac{(C)^2}{2n_{i1}}.$$

The value of $\bar{\rho}_{i2}$ could either be the same as that used in the experimental-group variance for study i or could be a value used for other control-group correlations, as in the example.

The variance of the difference $\hat{\Delta}_i$ computed using this *ad hoc* value to represent the sampling variance of the imputed control-group effect is

$$\text{var}(\hat{\Delta}_i) = \text{var}(d_{i1}) + \text{var}(\bar{\delta}_{i2}). \tag{21}$$

Analyses of the differences in mean changes are conducted by including the estimated differences in mean changes from no-control studies and their variances [computed via (20) or (21)] together with the complete data from the two-group studies.

Example. For this example retesting and maturation were assumed to have no effect, and the non-existent control-group mean changes were replaced with the value zero. The variance given in (20) was used in the computations.

The second pair of columns in Table 2 presents the $\hat{\Delta}_i$ values based on two-group comparisons (for the first five studies) and the estimates computed using $C = 0$ for the studies without control groups, and the variances. The analysis of these eight results is not independent of the analyses of the five studies discussed above. Ordinarily only the more comprehensive analysis including all studies would be conducted.

The analysis of the differences proceeded as described above, with an initial test of homogeneity. The homogeneity test statistic was 7.17, which was not significant when compared to percentage points of the chi-square distribution with 7 degrees of freedom. The weighted average difference was 0.35 with a standard error of 0.11.

Though the results appeared homogeneous, the differences (the $\hat{\Delta}_i$s) from the no-control and the two-group studies were compared. This analysis can be done via an analogue to planned comparisons (as shown here) or through a weighted analysis-of-variance procedure.

The estimate of Δ for the five two-group studies was 0.34 (SE = 0.16), whereas the estimate for the non-control studies, using the constant $C = 0$ for each control-group change was 0.37, also with a standard error of 0.16. A z test was used to compare the estimates of Δ for the two-group and no-control studies. The value of the z-test, $z = 0.13$, was not significant when compared to a table of standard normal deviates, indicating that the two-group and no-control studies did not differ significantly on treatment–control differences in motor-skill performance gains.

Agreement in the results of the two-group and no-control studies should be interpreted cautiously. Though the two kinds of study may truly not differ, they also may *appear* not to differ if zero is an inappropriate value for C. Consideration of the outcomes studied and the kinds of measures used can address this issue.

5.2. Imputing the control-group value of δ

The second method of including results of studies without control groups is to impute for each missing control-group change measure an estimate based on the results of existing control groups. This approach allows more flexibility in the replacement of missing values and also makes use of the information present in the results of the two-group studies, but it is more complex statistically.

Substituting an estimate rather than an arbitrary constant introduces an additional stochastic element into the estimation of Δ_i and its variance, as well as interrelationships between the new Δ estimates (denoted $\tilde{\Delta}_i$) and the data used to obtain the imputed estimates. Thus reviews using imputed control-group results require the use of generalized weighted least squares analyses (e.g. Hedges & Olkin, 1985; Raudenbush, Becker & Kalaian, 1988) or some other method of accounting for intercorrelations.

Imputed estimates and their variances. The new estimate of Δ is the difference between the imputed control-group value, denoted $\tilde{\delta}_{i2}$ and the ith experimental standardized mean-change measure. The estimate of Δ_i is

$$\tilde{\Delta}_i = d_{i1} - \tilde{\delta}_{i2}.$$

This approach allows a different value to be imputed for each no-control study, providing maximum flexibility in modelling between-study differences. (For example, $\tilde{\delta}_{i2}$ values could be estimated using a regression model computed for existing control-group data, using as predictor variables any study features observed for all studies.)

The variance of the new estimate is

$$\text{var}(\tilde{\Delta}_i) = \text{var}(d_{i1}) + \text{var}(\tilde{\delta}_{i2}).$$

Covariances of the imputed values. Imputation creates dependencies between each new Δ estimate and the data used in the imputation, as well as among the new estimates if several studies lack control groups.

The specific nature of the dependency depends upon the form of the imputed statistic. The general form of the covariance between the estimate of Δ from study i (which has both a treatment and a control group) and that from study m (which has

only a treatment group) is

$$\begin{aligned}\mathrm{cov}(\tilde{\Delta}_i, \tilde{\Delta}_m) &= \mathrm{cov}(d_{i1} - d_{i2}, d_{m1} - \tilde{\delta}_{m2}) \\ &= \mathrm{cov}(d_{i1}, d_{m1}) - \mathrm{cov}(d_{i1}, \tilde{\delta}_{m2}) - \\ &\quad \mathrm{cov}(d_{i2}, d_{m1}) + \mathrm{cov}(d_{i2}, \tilde{\delta}_{m2}) \\ &= \mathrm{cov}(d_{i2}, \tilde{\delta}_{m2}).\end{aligned}$$

The covariance between the differences depends only on the actual and imputed control-group change measures. The covariance between two estimates both involving imputed values for δ is

$$\begin{aligned}\mathrm{cov}(\tilde{\Delta}_i, \tilde{\Delta}_m) &= \mathrm{cov}(d_{i1} - \tilde{\delta}_{i2}, d_{m1} - \tilde{\delta}_{m2}) \\ &= \mathrm{cov}(\tilde{\delta}_{i2}, \tilde{\delta}_{m2}).\end{aligned}$$

The control-group mean-change values to be imputed are computed from the existing control-group mean changes. The computations are patterned after those described by Hedges (1983*b*), but utilize the data from only the control samples of the two-group studies. The imputation would include two (or more) steps, the first comprising analysis of existing control-group results (to obtain values of $\tilde{\delta}_{i2}$), and the consequent steps involving refinement of the imputation procedure and computation of the Δ estimates using the $\tilde{\delta}_{i2}$ values.

Adequacy of imputed estimates. When values are imputed for non-existent control-group results it is useful to ask how reasonable they are, or, how well they represent the results that would have been obtained had control groups been used. One way to examine the adequacy of an imputation scheme is to ask how well values obtained by that imputation scheme would represent the results of the existing control groups (i.e. the results on which the imputation scheme is based). In fact, it is impossible to determine how representative the imputed values really are since the data they replace are not missing (i.e. existent but unreported) but rather they do not exist. Thus the best obtainable information relies on the existent results.

Using existing control data to assess the adequacy of the imputation scheme requires the assumption that the no-control studies are similar (on all important variables) to the two-group studies. If the imputation plan produces reasonable values for the data from which it is derived, then it may also produce good imputed values for a similar set of studies without control groups. However, if a tentative imputation scheme does not give reasonable values for the two-group studies, there is little reason to believe that it will produce good values for a set of similar no-control studies. Values provided by an imputation scheme which is associated with specification error could provide reasonable estimates of non-existent results if the no-control studies differ from the two-group studies.

Assessments of model adequacy enable the reviewer to refine an imputation scheme before analysing the imputed values. If a tentative, simple scheme has significant specification error, more complicated imputation rules may improve the specification. For example, a reviewer planning to substitute a single value for all non-existent control-group results would examine whether that value adequately characterized or

'fit' all existing control groups. The reviewer would not want to substitute a single value for the control-group standardized mean change if two identifiable kinds of control group had different population values for δ.

More complex imputation schemes (e.g. substituting values according to a regression based on study features) would be evaluated by studying the specification of the relevant statistical models for the control results.

The analyses for model specification described in Hedges (1983*b*) for various kinds of hypothesized population models can be used to examine the adequacy of an imputation scheme. However, analyses of model specification are complicated by the dependence of the control-group standardized mean changes (the δ_{i2}s) and the experimental–control differences (the Δ_is).

Because the δ_{i2}s and the Δ_is are related by definition, their estimates are also related. This dependence leads to increased levels of Type I error when both sets of estimates are analysed. One approach is to reduce the preset significance level by using $\alpha/2$ rather than α for both specification analyses. There is also an argument for not reducing the overall α level, however. When assessing model specification the research hypothesis is confirmed by acceptance of the null model rather than by rejection. Thus higher levels of Type I error act to make the tests more conservative. The α level was not reduced in the example below.

Example. The simplest scheme for imputing the non-existent control-group results is to substitute the average (denoted $d_{.2}$) of the standardized changes for the existing control-groups. The weighted average control-group standardized mean change for the five two-group studies in Table 1 was 0.36 with a standard error of 0.10. Values of $\tilde{\Delta}$ imputed using the mean control-group result were thus smaller than the Δ estimate obtained using the constant zero. The estimates for the three no-control studies were 0.26 (Bissonette, 1965), -0.11 (Standridge, 1971), and -0.06 (Tufts, 1963). For the example data, estimates of change based on $C = 0$ were more liberal than those based on the control-group mean. However,this will not always be the case.

The variance of $\tilde{\Delta}_i$ with the average control change imputed was the sum of the variance of the ith experimental mean-change measure and the variance of the average control change. Specifically,

$$\begin{aligned} \text{var}(\tilde{\Delta}_i) &= \text{var}(d_{i1}) + \text{var}(d_{.2}) \\ &= \text{var}(d_{i1}) + 0.0104. \end{aligned}$$

The variances for the three studies using imputed control-group results are the sixth through eighth diagonal entries in the variance–covariance matrix in Table 3.

Substitution of the mean $d_{.2}$ for non-existent control-group results requires the assumption that the average is a reasonable characterization of the results of all control groups. This assumption was checked by testing the homogeneity of the control-group changes. The existing control-group effects were not homogeneous, with a chi-square value of 23.45 (d.f. = 4, $p < 0.05$) based on the five values of d_{i2} in Table 1.

This specification test suggests that the mean control-group change is not a good estimate of change for all of the existing control groups, and thus may not adequately characterize the no-control studies. However, because it is the simplest value which

Table 3. Variance–covariance matrix involving imputed standardized mean changes

	Study							
	1	2	3	4	5	6	7	8
1	0.12	0.00	0.00	0.00	0.00	0.01	0.01	0.01
2	0.00	0.12	0.00	0.00	0.00	0.01	0.01	0.01
3	0.00	0.00	0.40	0.00	0.00	0.01	0.01	0.01
4	0.00	0.00	0.00	0.15	0.00	0.01	0.01	0.01
5	0.00	0.00	0.00	0.00	0.09	0.01	0.01	0.01
6	0.01	0.01	0.01	0.01	0.01	0.10	0.01	0.01
7	0.01	0.01	0.01	0.01	0.01	0.01	0.09	0.01
8	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.07

may be substituted, the mean is used in this illustration. In practice a reviewer would try to find an imputation scheme based on a model that is more well-specified for the existing data.

Imputation of the average standardized mean change created dependencies involving both the new estimates and the Δ estimates from two-group studies. The covariance between the estimates of Δ from studies i (with two groups) and m (with only a treatment group) was

$$\begin{aligned}\mathrm{cov}(\hat{\Delta}_i, \tilde{\Delta}_m) &= \mathrm{cov}(d_{i1} - d_{i2}, d_{m1} - d_{.2}) \\ &= \mathrm{cov}(d_{i2}, d_{.2}) = \mathrm{var}(d_{.2}) = 0.0104,\end{aligned}$$

and the covariances among the three no-control results were also equal to $\mathrm{var}(d_{.2})$ or 0.0104. Table 3 shows the covariances for the eight studies.

A generalized weighted least squares regression model was estimated, using as predictors a vector of ones and an indicator of whether each study had a real control group. The estimated overall average difference in change $\hat{\Delta}$ was zero (SE = 0.19), and the coefficient representing whether the study had a control group was 0.33 (SE = 0.20). This value implies that the estimated experimental–control difference in change was one-third of a standard deviation larger for two-group studies than for no-control studies (even with imputed control-group results).

Neither regression coefficient was statistically significant at the 0.05 level ($H_R = 4.73$, d.f. $= 2, p < 0.10$), nor was any significant variation left unexplained by the regression ($H_E = 7.16$, d.f. $= 6, p > 0.25$). However, the total sum of squares for these data (i.e. $H_R + H_E$) was also not significant (its value was 11.89, d.f. $= 8$, $p > 0.10$), indicating that the hypothesis that all of the Δ values for this data set were equal *and* equal to zero could not be rejected.

The estimate of the treatment effect based on imputed values was much smaller than that obtained using zero for the non-existent control results, and suggested that mental practice produced no significant advantage beyond what would be gained with repeated measurement. However, these results must be cautiously interpreted because of the misspecification of the imputed value (the control mean change) for the data from which it was derived.

6. Solutions to other problems suggested by standardized mean-change measures

Use of measures of standardized mean change suggests solutions to at least two other problems in meta-analysis. Specifically, the use of standardized mean changes simplifies the synthesis of results from studies with multiple experimental and control groups, and aids in the study of differential treatment effects on multivariate outcomes.

When several experimental groups are compared to a single control group, multiple experimental-versus-control effect sizes computed using Glass's formula are intercorrelated in a complex way. Some meta-analysts have ignored this intercorrelation and treated multiple effect sizes based on a single control as independent. Other reviewers have combined data from the different treatment groups or omitted data from all but one treatment group in order to obtain independent estimates.

Intercorrelations between pairs of differences in mean changes based on a single control group are much simpler than for Glass's effect size, and can be treated appropriately in a meta-analysis (e.g. Hedges & Olkin, 1985; Raudenbush *et al.*, 1988; Rosenthal & Rubin, 1986).

Standardized measures of mean change are also useful for analysing multivariate outcomes. Comparisons not available by analysing Glass's effect sizes can be made. For example, the reviewer can model different gains due to maturation or retesting for different outcomes, as well as differential effects of treatments on outcomes.

The extensions of methods described above which allow inclusion of no-control studies can be generalized to be used with studies having multiple groups and multiple outcomes, as well.

Summary

Standardized measures of mean change can provide an alternative approach to the meta-analysis of data from studies using pre-test–post-test data-collection designs. Standardized mean-change measures can be analysed using fixed-effects, random-effects, or mixed models, depending upon the assumptions the researcher is willing to make concerning behaviour of and relationships among population mean-change parameters. These measures have reasonable statistical properties and their use may simplify the synthesis of results from studies with multiple treatment groups and multivariate outcomes. However, analyses of standardized mean changes (which rely on two-wave data-collection designs) should not be chosen over more informative and powerful longitudinal designs in the conduct of future primary research. Questions of how to synthesize data from these more complex designs provide opportunities for further inquiry.

Acknowledgements

Thanks go to Larry V. Hedges, Stephen W. Raudenbush, Harvey Goldstein, and an anonymous reviewer for feedback on earlier versions of this paper. This work was conducted while the author held a National Academy of Education Spencer Fellowship.

References

Bissonette, R. (1965). The relative effects of mental practice upon the learning of two gross motor skills. Unpublished master's thesis, Springfield College.

Brogan, D. R. & Kutner, M. H. (1980). Comparative analyses of pretest–posttest research designs. *The American Statistician*, **34**, 229–232.

Bryk, A. S., Raudenbush, S. W., Seltzer, M. & Congdon, R. T. (1986). *An Introduction to HLM: Computer Program and Users' Guide*. (Available from A. S. Bryk, Department of Education, University of Chicago, 5835 South Kimbark, Chicago, IL 60637.)

Campbell, D. P. & Stanley, J. C. (1963). *Experimental and Quasi-experimental Designs for Research*. Chicago: Rand-McNally.

Clark, L. V. (1960). Effect of mental practice on the development of a certain motor skill. *Research Quarterly*, **31**, 560–569.

DerSimonian, R. & Laird, N. M. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis. *Harvard Educational Review*, **53**, 1–15.

Feltz, D. & Landers, D. M. (1983). The effects of mental-practice on motor-skill learning and performance: A meta-analysis. *Journal of Sports Psychology*, **5**, 25–57.

Feltz, D., Landers, D. M. & Becker, B. J. (1986). A revised meta-analysis of the mental-practice literature. Paper commissioned by the National Research Council of the National Academy of Sciences.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, **5**, 3–8.

Glass, G. V., McGraw, B. & Smith, M. L. (1981). *Meta-analysis in Social Research*. Beverly Hills: Sage.

Goldstein, H. (1979). *The Design and Analysis of Longitudinal Studies*. London: Academic Press.

Hedges, L. V. (1981). Distribution theory for Glass's effect size and related estimators. *Journal of Educational Statistics*, **6**, 107–128.

Hedges, L. V. (1983*a*). A random effects model for effect sizes. *Psychological Bulletin*, **93**, 388–395.

Hedges, L. V. (1983*b*). Combining independent estimators in research synthesis. *British Journal of Mathematical and Statistical Psychology*, **36**, 123–131.

Hedges, L. V. & Olkin, I. (1985). *Statistical Methods for Meta-analysis*. New York: Academic Press.

Johnson, N. L. & Kotz, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions – 2*. New York: Wiley.

Kovar, S. V. (1969). The relative effects of physical, mental, and combined mental–physical practice in the acquisition of a motor skill. Unpublished master's thesis, University of Illinois.

McGaw, B. & Glass, G. V. (1980). Choice of the metric for effect size in meta-analysis. *American Educational Research Journal*, **17**, 325–337.

Perry, H. M. (1939). The relative efficiency of actual and imaginary practice in five selected tasks. *Archives of Psychology*, **34**, 5–75.

Raudenbush, S. W., Becker, B. J. & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin*, **103**, 117–120.

Raudenbush, S. W. & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics*, **10**, 75–98.

Rawlings, E. I. & Rawlings, I. L. (1974). Rotary pursuit tracking following mental rehearsal as a function of voluntary control of mental imagery. *Perceptual and Motor Skills*, **38**, 302.

Rosenthal, R. & Rubin, D. B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin*, **92**, 500–504.

Rosenthal, R. & Rubin, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin*, **99**, 400–406.

Ryan, D. E. & Simons, J. (1982). Efficacy of mental imagery in enhancing mental rehearsal of motor skills. *Journal of Sport Psychology*, **4**, 41–51.

Standridge, J. O. (1971). The effect of physical, mental, and mental–physical practice in learning the whip kick. Unpublished master's thesis, University of Tennessee.

Tufts, S. A. (1963). The effects of mental practice and physical practice on the scores of intermediate bowlers. Unpublished master's thesis, University of North Carolina at Greensboro.

Whitehill, M. P. (1964). The effects of variations of mental practice on learning a motor skill. Unpublished master's thesis, University of Oregon.

Received 19 March 1987; revised version received 25 November 1987