Bulletin of the Psychonomic Society
1988, 26 (1), 51-54

A method for analyzing retrospective pretest/posttest designs: I. Theory

PIETER KOELE and JOHAN HOOGSTRATEN
University of Amsterdam, Amsterdam, The Netherlands

To control for response-shift bias in pretest/posttest designs, Howard and his collaborators proposed the use of retrospective pretests, the so-called thentests. If a response-shift bias is present, thentest/posttest difference scores will yield better estimates of a treatment effect than will the usual pretest/posttest difference scores. In this article a method is presented that is aimed at detecting response-shift bias in retrospective pretest/posttest designs. This method, essentially a hierarchical model fitting procedure, is easy to apply, prevents capitalization on chance, and makes optimal use of information supplied by the control group.

Studies using self-reports in a pretest/posttest design may be contaminated by the occurrence of a response-shift bias. As a control for this kind of contamination, the use of a retrospective pretest (the so-called thentest) has been recommended (Howard & Dailey, 1979). Studies investigating the potential validity of the retrospective pretest/posttest design (e.g., Hoogstraten, 1982, 1985; Howard, 1980; Howard, Ralph, Gulanik, Maxwell, Nance, & Gerber, 1979) in general have shown positive results. Bray, Maxwell, and Howard (1984) therefore asserted that the proper way to obtain unbiased estimates of a treatment effect in a retrospective pretest/posttest design is to analyze posttest minus thentest difference scores. This recommendation, however, takes for granted the presence of a response-shift bias. Because experimental results are hardly ever as neat and clear as may be hoped for, the presence of response-shift bias cannot always be clearly established, and the experimenter is consequently more or less at liberty to use either pretest/posttest or thentest/posttest difference scores for estimating the treatment effect. A bias in favor of the experimenter's hypothesis is then most likely to occur. To prevent this bias, a test on the presence of a response-shift bias should be executed before treatment effects are estimated.

Bray et al. (1984) further stated that the existing literature strongly suggests that posttest minus thentest difference scores are more valid than posttest minus pretest difference scores. Typically, an objective criterion is used in studies designed to assess the relative validity of posttest/thentest versus posttest/pretest difference scores. The accumulated empirical evidence on this issue is not equivocal. Howard, Millham, Slaten, and O'Donnell (1981) stated that in general thentest/posttest scores are more highly correlated with objective measures of change than are pretest/posttest scores. However, some recent studies have reported correlations between self-reported change and objective measured change that are not significant and are often, in fact, close to zero (Hoogstraten, 1982, 1985; Sprangers & Hoogstraten, 1987). This leaves the empirical issue of the relative validity of pretest/posttest and posttest/thentest scores unsettled.

The purpose of the present study is to present a method for analyzing retrospective pretest/posttest designs. In this method the detection of a response-shift bias is the central issue. If this bias can be demonstrated, the use of thentest/posttest difference scores for estimating treatment effects is justified. If not, pretest/posttest difference scores should be used.

DESIGN

The basic retrospective pretest/posttest design is given in Table 1. Subjects are randomly assigned to the experimental and the control conditions, and mean scores on the dependent variable Y (usually self-report) are of primary interest. A response-shift bias "is conceived as effects produced by the treatment which results in a shifting of subjects' perceptions of themselves" (Bray et al., 1984, p. 783). In the experimental condition this will result in a mean score on the thentest that differs from the mean on the pretest. Because the bias is dependent on the presence of the experimental treatment, it should not occur in the control condition. Thus, a response-shift bias may be present when

$$|\overline{Y}_E(\text{then}) - \overline{Y}_E(\text{pre})| > |\overline{Y}_c(\text{then}) - \overline{Y}_c(\text{pre})| \quad (1)$$

This formalization is less strict than the definition of a response-shift bias, but it should be kept in mind that sample means fluctuate, and Inequality 1 therefore represents a necessary but not a sufficient condition for the existence of a response-shift bias. A sufficient "proof" (within the limits of statistical inference) can be given only by comparing the fit of theoretical models with and without response-shift bias on the experimental results.

The authors are indebted to Mirjam Sprangers for her stimulating comments. Address correspondence to Pieter Koele, Methodology Unit, University of Amsterdam, Weesperplein 8, 1018 XA Amsterdam, The Netherlands.

Table 1
Basic Retrospective Pretest/Posttest Design

	Pretest	Treatment	Posttest	Thentest
Experimental Condition	$\overline{Y}_E$(pre)	×	$\overline{Y}_E$(post)	$\overline{Y}_E$(then)
Control Condition	$\overline{Y}_c$(pre)	–	$\overline{Y}_c$(post)	$\overline{Y}_c$(then)

The retrospective pretest/posttest design can be seen as a two-way factorial design with repeated measures on one factor. In the most simple design, there is a between-subjects factor groups with two levels (experimental and control) and a within-subjects factor measurement with three levels (pre, post, and then). In this design a significant groups × measurement interaction effect indicates the presence of a treatment effect and/or response-shift bias. If such an interaction effect has been obtained, the analysis can proceed, focusing on the scores in the experimental group. The scores in the control group are used to yield estimates of systematic effects due to history, maturation, testing, and so forth. The scores in the experimental group are corrected for these effects. The corrected scores are analyzed by a one-way analysis of variance with repeated measures on the factor measurement (pre, post, and then). Four different planned comparisons (contrasts) between means are specified, which correspond with different configurations of treatment effects and/or response-shift bias. The contrast that explains the highest significant percentage of variance between the three cell means is finally selected as giving the most optimal description of the experimental results.

RESPONSE-SHIFT BIAS MODELS

On population level, differences between μ_c(pre), μ_c(post), μ_c(then) and their grand mean $\mu_{c.}$ reflect systematic differences that are not a result of the experimental manipulation. The purpose of a randomized pretest/posttest design is to evaluate differences between means in the experimental group relative to differences between corresponding means in the control group. It is, therefore, legitimate to subtract the estimated effects $\hat{\gamma}$(pre) = $\overline{Y}_c$(pre) − $\overline{Y}_{c.}$, $\hat{\gamma}$ (post) = $\overline{Y}_c$(post) − $\overline{Y}_{c.}$, and $\hat{\gamma}$(then) = $\overline{Y}_c$(then) − $\overline{Y}_{c.}$ from the corresponding scores in the experimental group, yielding the three corrected cell means $\overline{Y}'_E$(pre) = $\overline{Y}_E$(pre) −$\hat{\gamma}$ (pre), etc. These corrected scores and the corrected cell means form the basis of the subsequent analysis.

Let us assume that in a particular experiment the experimenter expects a beneficial treatment effect and a response-shift bias that will lower scores on the thentest, compared with scores on the pretest. Given these expectations, the following models can be specified, depending on the absence or presence of a treatment effect and/or response-shift bias.

1. *Treatment effect present, no response-shift bias.* In this model $\overline{Y}'_E$(pre) equals $\overline{Y}'_E$(then), and $\overline{Y}'_E$(post) is higher than the other two. Graph 1 in Figure 1 illustrates this model.

2. *Response-shift bias present, no treatment effect.* This model is probably the most tricky one; there seems to be a treatment effect in an unexpected direction (in terms of pretest/posttest difference scores), but it really is only a response-shift bias (Graph 2 of Figure 1).

3. *Treatment effect and response-shift bias present.* This model is illustrated in Graph 3 of Figure 1. Notice that in this case the pretest/posttest difference scores give an underestimate of the treatment effect.

4. *Treatment effect present but hidden, response-shift bias also present.* A comparison of pretest/posttest difference scores would lead to the conclusion that there is no treatment effect. The response-shift bias overshadows the treatment effect (Graph 4 of Figure 1).

These four models represent predictions about the corrected cell means in the experimental group in a retrospective pretest/posttest design, given the presence or absence of a treatment effect and/or a response-shift bias. There exists a hierarchy in these models. Models 1 and 2 are the simple ones, assuming only either a treatment effect or a response-shift bias. Models 3 and 4 are more complex, because they assume the existence of both a treatment effect and a response-shift bias. The order in which these models should be fitted to the experimental results

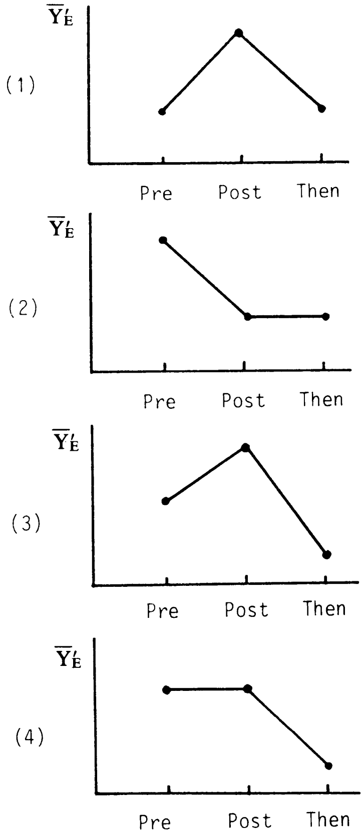

Figure 1. Response-shift models. (See text for details.)

is dictated by their degree of complexity. First, the so-called null model is tested. This model represents the absence of both a treatment effect and a response-shift bias. If the null model can be rejected, the scores in the experimental group are corrected for the systematic effects between means in the control group. On these corrected data, Models 1 and 2 are fitted. When these two models do not give a good fit, Models 3 and 4 are fitted. When these models also do not fit nicely, two explanations are possible. Either the experimental results are simply a mess and no conclusion can be drawn about treatment effect and/or response-shift bias, or the expectations of the experimenter about the direction of treatment-effect and response-shift bias were mistaken, and other models should have been specified and fitted. A priori, the latter explanation is the most unlikely one, of course.

In the next section the particulars of the model fitting procedure will be given. It should be clear by now that this procedure determines for each model to be fitted how much of the between-cell-means variance it can explain. To this purpose the method of planned comparisons (or contrasts) between cell means is used.

THE MODEL FITTING PROCEDURE

Fitting the Null Model

The null model reflects the absence of both a treatment effect and a response-shift bias. It is rejected if the groups × measurement interaction of the 2×3 ANOVA with repeated measures on measurement reaches significance.

Fitting Models 1 and 2

Before proceeding after the rejection of the null model, the scores in the experimental group must be corrected along the lines sketched above.

Next, a one-way analysis of variance with repeated measures on measurement is executed, using these corrected scores. There is no loss of degrees of freedom involved, because the correction is based on estimates obtained on data independent of those in the experimental group. Furthermore, the correction does not influence the variance within cells. The method used for fitting Models 1 up to 4 on these corrected cell means is essentially similar to the method of testing contrasts or testing for trend (see Winer, 1971; or Ferguson, 1976). For each of the models in Figure 1, the cell means are predicted by assigning a numerical value (*coefficient*) to them, representing their relative magnitude. For the sake of arithmetical convenience, these coefficients must add to zero per model; otherwise, there are no restrictions. In Table 2 a possible display of four sets of coefficients is given. The models are not orthogonal. This means that tests on their goodness of fit are interdependent; in a later section this issue is discussed.

In order to calculate the proportion of variance explained by a model, one calculates the Pearson correlation between the coefficients of this model (the predicted cell means) and the corrected cell means, and one squares this correlation. The next step is to evaluate these proportions of explained variance, in order to decide whether the model should be accepted or rejected. The proper way to do this is to test whether the amount of variance *not* explained by a particular model differs significantly from zero. If a model fails to pass this test, there is evidently room for improvement, and other models may be tried. The amount of variance to be explained is *SS*(measurement), that is, the sums of squares associated with the main effect of measurement. This *SS*-term has 2 degrees of freedom. Because any model has 1 degree of freedom, the amount of variance not explained by a particular model has 1 degree of freedom. If p_i is the proportion of variance explained by Model i, the amount of variance not explained by Model i equals $(1-p_i) \cdot SS(\text{measurement})$, and the statistic used for testing whether this amount differs significantly from zero is given by

Table 2
Coefficients for the Models of Figure 1

	Pre	Post	Then
Model 1	−½	1	−½
Model 2	1	−½	−½
Model 3	0	1	−1
Model 4	½	½	−1

$$F_i = \frac{[(1-p_i) \cdot SS(\text{measurement})]}{MS(\text{measurement} \times \text{subjects})}, \qquad (2)$$

following a $F(1,2(n-1))$ distribution, where n is the number of observations per cell (assuming equal cell frequencies).

If the tests are carried out for Models 1 and 2, several results are possible. If the test in Equation 2 in both cases reaches significance, both models are rejected and Models 3 and 4 can be fitted. If one of the tests reaches significance and the other does not, the latter model must be accepted. If none of the tests reaches significance, the model with the largest amount of explained variance should be accepted.

Fitting Models 3 and 4

The procedure for these models follows the pattern laid out above. The result is either the selection of a best fitting model or the sad conclusion that none of the specified models gives an adequate description of the experimental results. When Model 2, 3, or 4 is selected, a response-shift bias has been demonstrated, and posttest/thentest difference scores should be used for estimating treatment effects.

An application of the procedure, reanalyzing data from an earlier study, is presented in Hoogstraten and Koele (in press).

DISCUSSION

The model fitting procedure described above has two important properties. First, the procedure is hierarchical in nature, leading from a global test to specific tests, thus avoiding the problem of capitalization on chance. Second, the procedure uses the information supplied by the control group in an optimal way. Given both the logic underlying the use of a control group in pretest/posttest designs and the linear, additive model underlying analysis of variance techniques, estimating systematic nonexperimental effects in a control group and subtracting them from scores in the experimental group is a natural and perfectly sound thing to do. Because of these two reasons, we consider the model fitting procedure to be the proper way for analyzing retrospective pretest/posttest designs.

A refinement in the procedure is possible by taking into account the covariance of the models in the model fitting procedure, in such a way that each model is fitted on cell-means from which the common variance with some or all other models is removed. This is done by calculating the relevant semipartial correlations between model predictions and cell-means, with other model predictions partialed out of the cell means. We feel, however, that this refinement is perhaps a bit too sophisticated and that it only complicates the evaluation of the goodness of fit of the models without giving clear benefits. The fact that the models do covariate is something that should be kept in mind, in order to resist the temptation to add percentages of explained variance: It is wonderful to explain 150% of variance, but it is a bit odd.

REFERENCES

Bray, J. H., Maxwell, S. E., & Howard, G. S. (1984). Methods of analysis with response-shift bias. *Educational & Psychological Measurement*, **44**, 781-804.

Ferguson, G. A. (1976). *Statistical analysis in psychology and education* (4th ed.). New York: McGraw-Hill.

Hoogstraten, J. (1982). The retrospective pretest in an educational training context. *Journal of Experimental Education*, **50**, 200-205.

Hoogstraten, J. (1985). Influence of objective measures on self-reports in a retrospective pretest/posttest design. *Journal of Experimental Education*, **53**, 207-210.

Hoogstraten, J., & Koele, P. (in press). A method for analyzing retrospective pretest/posttest designs: II. Application. *Bulletin of the Psychonomic Society*.

Howard, G. S. (1980). Response-shift bias: A problem in measuring change with self-report. *Evaluation Review*, **4**, 93-106.

Howard, G. S., & Dailey, P. R. (1979). Response-shift bias: A source of contamination of self-report measures. *Journal of Applied Psychology*, **64**, 144-150.

Howard, G. S., Millham, J., Slaten, S., & O'Donnell, L. (1981). Influence of subject response style effects on retrospective measures. *Applied Psychological Measurement*, **5**, 89-100.

Howard, G. S., Ralph, K. M., Gulanick, N. A., Maxwell, S. E., Nance, S. W., & Gerber, S. K. (1979). Internal validity in pretest-posttest self-report evaluations and a re-evaluation of retrospective pretests. *Applied Psychological Measurement*, **3**, 1-23.

Sprangers, M., & Hoogstraten, J. (1987). Response style effects, response-shift bias, and a bogus pipeline. *Psychological Reports*, **61**, 579-585.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd Ed.). New York: McGraw-Hill.

(Manuscript received for publication June 8, 1987.)