Review of Educational Research
Summer 1989, Vol. 59, No. 2, pp. 215–228

Hawthorne Control Procedures in Educational Experiments: A Reconsideration of Their Use and Effectiveness

John G. Adair, Donald Sharpe, and Cam-Loi Huynh
University of Manitoba

This article reports on a descriptive analysis of research practices and a meta-analysis of effect sizes associated with control groups employed to address Hawthorne effects in educational experiments. The descriptive analysis of 86 studies revealed a diversity of practices designed to control one of three artifact variables: special attention, activity related to the experimental task, or awareness of participation in an experiment. The meta-analysis provided no evidence for a Hawthorne versus no-treatment control difference. Moreover, a detailed analysis of these studies by their control procedure, and subsequently by other moderator variables, also revealed no systematic trends to suggest a specific artifact source. A within-study analysis of the pattern of treatment/Hawthorne/control group effect sizes suggested that the artifact controls were of limited utility. Increased research is urged into other artifacts and alternative control procedures that the educational researcher should consider.

Years ago Desmond Cook (1962) raised the specter of the Hawthorne effect, the methodological contaminant first observed in the industrial experiments known as the Hawthorne studies (Roethlisberger & Dickson, 1939), as an artifact that might contaminate experiments in education. His subsequent research and unpublished conclusion (Cook, 1967) that there was a lack of understanding of the Hawthorne effect and that it might be less contaminating than at first thought, had far less impact than his earlier pronouncement. Other studies and reviews with similar conclusions (Diamond, 1974; Schneiderman, 1977), also remain unpublished. On the other hand, Campbell and Stanley (1963) had compellingly argued for educational researchers to control for "reactive arrangements," and Bracht and Glass' (1968) subsequent listing of the Hawthorne effect as a threat to internal validity indelibly inscribed the artifact in experimental methods texts.

A recent review (Adair, 1984) has shown that the confusion surrounding the artifact has persisted, in large part because the source of the Hawthorne effect was never clearly identified. Dickson and Roethlisberger (1966), for example, attributed it to a combination of 17 possible mediators. Examination of Hawthorne control practices of educational researchers indicates that three salient features of the original studies have been identified as the primary source of the methodological

We want to thank Philip Abrami, Sylvia d'Apollonia, and two anonymous reviewers for their many helpful comments. A quite different, preliminary version of this article was presented at the annual meeting of the American Educational Research Association, Washington, D.C., April 1987. The research was supported by a grant to the first author from the Social Sciences and Humanities Research Council of Canada.

artifact: the special attention subjects received from their observers and supervisors; awareness of their participation in an experiment; and the novelty or unique features of the experimental activity (Adair, 1984). The control procedures associated with each artifact source were also found to vary, from special groups treated in various ways to merely waiting for the Hawthorne effects to dissipate over time. Although each of these procedures has some reference in experimental methods textbooks (Adair, 1981) to justify its appropriateness as a Hawthorne control, the "true" artifact source and the appropriate control procedure remain unclear.

The Hawthorne effect has not been satisfactorily defined. Moreover, classification of Hawthorne control procedures by primary source—that is, special attention, awareness, or novelty of the experimental activity—has not been examined by methodologists. The typical procedure is to provide a control group with some special treatment beyond that which a waiting-list control group would receive in order to assess the effect of the artifact. Various design strategies and "placebo-like" activities have been employed to address variables identified as Hawthorne artifacts. By accumulating, coding, and categorizing these practices, the range and patterns of Hawthorne control groups might be systematically described and examined. Cook (1967) attempted such a review but was hampered by the paucity of studies that had employed Hawthorne controls to that time. Due to the continued widespread use of Hawthorne controls in the absence of some standard definition, there is a need to assess the procedures of studies that explicitly use the term Hawthorne. In addition, a meta-analysis of the magnitude and the homogeneity of effect sizes associated with each control practice, and the pattern of treatment/Hawthorne/no-treatment-control effects within each study permits an objective determination of the effectiveness and appropriateness of each classified procedure.

Classification of Research Practices

Method

Source of Studies. An attempt was made to identify all educational studies that employed control groups labeled "Hawthorne" or control groups by some other name for which the stated rationale was the manipulation or control of "Hawthorne effects." Because we were interested in only those effects and control groups that educational researchers defined as Hawthorne, studies that employed comparable procedures with another label or that addressed artifact without using a specially treated control group were not included. Studies were accumulated from previous Hawthorne reviews (Adair, 1984; Cook, 1967; Diamond, 1974), and from full-text searches for the terms "Hawthorne" and "placebo" in ERIC, PsychInfo, and the on-line version of *Dissertation Abstracts International* databases. Additional ERIC searches were conducted on such broad terms as "experimental group" and "control group."

A total of 86 studies were identified for review.[1] Twenty-nine were published journal articles, 32 were ERIC documents or unpublished papers, and 25 were dissertations. Because of the extensive searches, the database may be regarded as representative of control groups reported by educational researchers that address Hawthorne effects.

Coding. Studies were coded on each of 18 separate variables, including (a) rationale for Hawthorne control as stated by study author(s) (control for attention,

activity, novelty, Hawthorne effect, interaction, motivation, and awareness); (b) control procedure (match attention, activity, or awareness of experiment participation, or some combination of these procedures); (c) Hawthorne control activity (minimal contact, on-going activity, substitute activity, treatment resemblance, or treatment element); (d) form of publication (dissertation, journal article, or ERIC report); (e) publication date; (f) topic classification (curriculum-related topics, classroom behavior, concept training, learning disabilities, mental retardation, Hawthorne effects, reading, personality and mental health, teaching methods, teacher performance-related factors, and other); (g) subject characteristics (age and special characteristics); (h) treatment administration (number of groups and mode of treatment administration); (i) no-treatment controls (whether employed or not); (j) subject sampling (whether random or otherwise); and (k) blind controls (whether any other special artifact control procedures were also employed). Other variables, such as experimenter characteristics, the length and nature of the treatment, outcome type and reactivity, and subjects' phenomenological states, were rated but not discussed because they did not reveal any systematic relationship to the control procedures. All codings were initially completed by the second author and an honors psychology student. Agreement between the two sets of ratings was generally quite high, with disagreements resolved by an independent reading of the article by the first author.

Results

The range and frequency of Hawthorne control procedures among the 86 studies are indicated in Table 1. In all studies the Hawthorne group was matched with the experimental group on one or more variables that reflected the investigator's artifact concern. Hawthorne control group subjects were most often given an activity to equate them with the experimental group in terms of time, effort, interest, or novelty. For example, Hawthorne subjects were given sedentary physical activities as a control for the physical training of experimental subjects (Chasey, Swartz, & Chasey, 1974) or control group lessons of the same duration and from the same source as those for the treatment group (Kalechstein, Kalechstein, & Docter, 1981). A second common practice was to tell Hawthorne controls that they were in an experiment, to match their awareness with that of the treatment group subjects (e.g., Campbell, 1978; Higgins & Rusch, 1965). A third practice was to equate the special attention that all subjects experienced, for example, by matching Hawthorne and treatment subjects on both the frequency and quality of experimenter-subject interaction (Herman, 1982).

To compound the diversity of procedures, these three manipulations occurred in various combinations, and occasionally with a fourth manipulation, namely special instructions or conditions to enhance the motivation of control subjects. In addition, there were several infrequently used nonmatching procedures, such as adding a second no-treatment control group at a different location or time to compare with the control group established within the experimental context (Pella, Stanley, Wedemeyer, & Wittich, 1962).

To make the analyses meaningful, all control procedures were grouped into one of the dominant Hawthorne concerns. In the case of the more complex procedures, that meant placing into the attention or awareness category any study that had manipulated that variable, usually in addition to matching the level of activity of

TABLE 1
Hawthorne control group procedures

Procedure	Number of studies
Matching Hawthorne Group (H) with experimental group (T) on:	
A single variable (although other dimensions may be incidental and/or implied, it is the variable below that defines the H):	
1. Activity—to equalize time, effort, interest, novelty	34
2. Awareness of experiment participation—includes telling both T and H, or only H, that they are in an experiment	14
3. Attention—to equalize contact or interaction	10
Multiple manipulation and matching (more than one variable is manipulated or introduced into the H to equate it to the T or to generate a Hawthorne effect):	
1. Attention and activity	12
2. Awareness and activity	3
3. Awareness, activity, and motivation (any special instruction or activity that is given to the H to arouse their level of enthusiasm or expectation for success or to enhance their performance)	3
4. Attention, activity, and motivation	5
Special treatment administered to Hawthorne group (activity of H nonexistent or incidental; specific effort to interact with Hawthorne control subjects beyond equating control and treatment subjects)	2
Create second control group (C) so that earlier group becomes H	
No-treatment control established in another location or independent of experiment so that it can be compared with T and H, without the awareness of experiment participation (presumably a condi- the awareness of experiment participation (presumably a condition that makes H a "flawed" C)	2
After experiment, a second C established to compare as a no-treatment control with T and H (i.e., C established during experiment and presumably aware of experiment)	1

the experimental group. The activity-matching category was reserved exclusively for studies in which matching time, effort, or task novelty was the sole manipulation. In Table 2, the frequency of studies of each type of Hawthorne control procedure are presented according to whether a no-treatment control group was included and the type of control task activity.

A substantial number of researchers employed what might be called an irrelevant or substitute activity ($N = 28$) for their Hawthorne control. Presumably the intention was to engage the control group in an activity equivalent to the experimental activity in attention, time, effort, or novelty. In the largest number of studies ($N = 36$), the Hawthorne activity was designed to resemble the treatment in medium of stimulus presentation, form of activity, or general nature of stimulus materials. A smaller subset of studies ($N = 16$) included some substantive element of the treatment procedure to determine, by comparison, whether the full treatment had a significantly greater effect on subjects. A few studies employed only minimal contact or on-going activities ($N = 6$) as the form of Hawthorne task.

TABLE 2
Use of no-treatment controls and combinations of control group activity for each Hawthorne procedure[a]

Activity	No-treatment controls absent ($N = 32$)			No-treatment controls present ($N = 54$)			Total ($N = 86$)		
	Attn	Act	Awa	Attn	Act	Awa	Attn	Act	Awa
Irrelevant	10	3	1	6	7	1	16	10	2
Treatment resemblance	2	6	3	5	16	4	7	22	7
Treatment element	1	2	3	2	0	8	3	2	11
Other	1	0	0	2	0	3	3	0	3
Total	14	11	7	15	23	16	29	34	23

[a] Hawthorne procedures: Attn = attention; Act = activity; Awa = awareness; studies employing combinations of the variables have been collapsed into either Attn or Awa.

The combination of the type of activity and Hawthorne procedure revealed some interesting patterns. When special attention was the problem that the Hawthorne procedure was designed to address, most often the control group was given an irrelevant activity (55.2%). On the other hand, when mere awareness of experimentation was addressed by the Hawthorne procedure, control subjects were generally given a substantial element of the experimental treatment (47.8%) or an activity with treatment resemblance (30.4%). The overwhelming majority of studies that matched activity as the Hawthorne procedure employed a treatment resemblance activity (64.7%), although a significant subset of these studies (29.4%) matched effort, time at task, etc., on an irrelevant activity.

A serious shortcoming in all Hawthorne control studies was the failure to provide evidence or justification for the control procedure. The norm was to assume that the matching process was adequate or to describe the activity or procedure as if it were self-evident why this particular choice was appropriate. The placebo-like qualities of the Hawthorne control activity were documented by citation in only four studies.

In addition to these basic design features, there were other variables on which the studies systematically varied. For example, over half of the studies ($N = 44$) were conducted with child subjects. Cook (1967) and others have suggested that children might be less susceptible to Hawthorne effects than other subject populations. Within our sample, 62% of attention-manipulating studies, 47% of activity-matching studies, and 41% of studies employing awareness controls were conducted with elementary-school-aged child subjects. Of the studies in which awareness was manipulated, 65% used intact classes or other non-random sampling procedures. In contrast, 62% of experiments that controlled by matching activities and 61% of experiments that matched on attention used some form of random sampling.

The sample of studies was evenly distributed across a number of research topic areas: academic curricula ($N = 3$), school or classroom behavior ($N = 7$), concept training ($N = 9$), Hawthorne studies ($N = 14$), learning disabilities ($N = 8$), mental retardation ($N = 15$), personality and mental health ($N = 2$), reading ($N = 12$), teaching methods ($N = 11$), teacher performance-related factors ($N = 3$), and other educational ($N = 2$). This suggests that the concern with Hawthorne was general,

although there were differences between topic areas. For example, only 8% of the reading experiments assigned the Hawthorne label to a control group, whereas 60% of the retardation and 62.5% of the learning disability experiments included a labeled Hawthorne control group. Unfortunately, the limited frequency of studies within virtually all topic areas precluded further analyses of control group effectiveness by topic area.

Discussion

Because artifact controls must be tailored to each study, some diversity of procedures was to be expected. However, the fact that there should be "a spectrum of control conditions" (Runkel & McGrath, 1972, p. 226) does not account for the varied procedures labeled Hawthorne controls. The range of procedures is consistent with the confusion regarding the Hawthorne artifact and its appropriate controls. Definitions of Hawthorne were seldom explicitly stated and could only be inferred from the variable(s) emphasized in the control procedure. When offered, different definitions of the Hawthorne artifact were emphasized and seemed to suggest different control procedures.

Independent of their validity, three artifact concerns and accompanying control procedures emerged from the data: (a) *Special attention*: Control subjects (predominantly children, randomly assigned to conditions) were given some contact or interaction with the experimenter that was assumed to be equivalent to that given experimental subjects. Their task activity, however, generally bore no relationship to the treatment group activity. (b) *Awareness*: Control subjects (predominantly adults) were typically equated with experimental subjects on their knowledge of participation in an experiment. As a consequence, most awareness controls consisted of intact classes that received an activity substantially similar to the treatment. (c) *Activity matching*: Randomly assigned control subjects (both children and adults) were given an activity that resembled the experimental task in form or content, that matched it in novelty, or that required equal time and effort. Although there were numerous variations on each procedure, in every case the researcher claimed to be controlling for a Hawthorne effect.

Meta-Analysis

The meta-analytic phase of the research was undertaken to provide a quantitative assessment of the effect of each type of control practice. Of the original 86 studies, 32 studies were excluded from the meta-analysis because they did not include a no-treatment control group, and 15 studies were excluded because they failed to provide adequate statistical information. Studies were not excluded because of poor design or other quality considerations. The effective sample thus consisted of 39 studies on which effect sizes could be computed. These studies seemed to be representative of the 86 Hawthorne studies described above, with the exception that more recent studies tended *not* to employ a no-treatment control (NTC) group. Nonetheless, one-third of the studies in the meta-analysis were published in the most recent decade.

Method

An unbiased estimator of effect size, d (Hedges, 1981; Hedges & Olkin, 1985), was computed for both treatment and Hawthorne controls by subtracting the mean

for the no-treatment control group (NTC) from the appropriate mean in each case, dividing by the pooled standard deviation, and multiplying by Hedges' correction to remove the bias associated with small samples (Hedges, 1981). For those studies where means and standard deviations were unavailable, effect sizes were computed, where possible, by appropriate procedures as outlined by Glass, McGaw, and Smith (1981).

Results

Across-Study Analysis. Because of multiple dependent measures or multiple treatment groups in many studies, a total of 256 treatment/Hawthorne/NTC effect-size comparisons were computed. Six observations were classified as outliers and were discarded. The observations from each study were pooled, resulting in a sample of 38 effect sizes (two of the discarded outliers were observations from one study, with the result that the study was deleted).

A homogeneity test for independent effect sizes (Hedges & Olkin, 1985) indicated that the 38 Hawthorne/NTC effect sizes could be regarded as sharing a common effect-size distribution ($H_T(37) = 44.71$, $p > .05$). The average effect size obtained from the 38 observations was .33 for the treatment and .10 for Hawthorne. The corresponding weighted mean effect sizes were .20 and .01, respectively. Construction of 95% confidence intervals for the weighted mean effect sizes (Hedges & Olkin, 1985, p. 113) indicated that population effect sizes differed from zero for treatment (+.13 to +.28) but not for Hawthorne (−.07 to +.08). Substantially larger mean treatment and placebo effect sizes ($M = .93$ and .56, respectively) have been reported for psychotherapy outcomes (Smith, Glass, & Miller, 1980). Although educational procedures may have much less impact on subjects' responses than analogous psychotherapy procedures, this lesser treatment effect may also be due to the diversity of educational procedures that were grouped to arrive at a mean effect size. Such global comparisons of diverse sets of studies may be inappropriate (Hedges & Olkin, 1985); however, the absence of a significant Hawthorne effect is noteworthy in view of the literature arguing that such an effect can be reliably produced and needs to be guarded against.

In order to add to our understanding of the Hawthorne effect, and because Hedges (1982) suggests that analysis of moderator variables may proceed regardless of the outcome of the overall test of homogeneity, effect-size clusters were created on the basis of control procedure as the most theoretically relevant moderator variable. The number of studies classified as employing each control procedure, the unweighted and weighted mean effect sizes, standard deviations, and the 95% confidence intervals for the weighted mean effect sizes are all reported in Table 3. Weighted mean effect sizes are the product of each effect size weighted by the sample size of the study (Hedges & Olkin, 1985). An average sample size was calculated for those studies where effect sizes were derived from different sample sizes within the same study. The test of within-group homogeneity (H_w) indicated that the attention ($H_w(9) = 13.63$, $p > .05$), activity ($H_w(19) = 19.38$, $p > .05$), and awareness ($H_w(7) = 11.12$, $p > .05$) clusters were all homogeneous. The test of differences across control procedure clusters (H_B) was not significant ($H_B(2) = .56$, $p > .05$). Construction of 95% confidence intervals revealed that none of the mean effect sizes differed significantly from zero.

In addition to control procedure, nine other variables on which the studies had been coded, and for which meaningful clusters could be derived, were subjected to Hedges' analysis for moderator variables (Table 4). Based on the test of homogeneity within each cluster (H_W), all but two clusters created from partitioning of the 38 weighted effect sizes by moderator variables were judged homogeneous. As is evident from the statistically nonsignificant values for H_B, the weighted effect sizes did not differ across clusters for any of the moderator variables. Similarly, construction of 95% confidence intervals indicated that the weighted mean effect sizes associated with each cluster did not differ from zero.

Within-Study Analysis. Although the foregoing analyses differentiate the size of the Hawthorne effect associated with each manipulation, a "within-study" analysis of the pattern and magnitude of treatment/Hawthorne/NTC group effect sizes is informative with respect to the adequacy of each control procedure. Whether or not a Hawthorne procedure is satisfactory depends upon its position relative to both the treatment and NTC groups. The model of an appropriately conceived Hawthorne procedure derived from the literature is of a group that should manifest some greater effect than a control group, yet not so much of an effect that subjects perform comparably to those exposed to the experimental treatment. A "good" case is thus one in which the Hawthorne group falls between a significant treatment and an NTC group. In the "weak" case, the Hawthorne treatment is so inappropriately weak that these subjects perform essentially the same as NTC subjects do. The "nonsignificant" case represents the situation in which all comparisons are statistically nonsignificant. "Rare" cases also yield uninterpretable Hawthorne results because it is difficult to know the meaning of a significant Hawthorne result when the matching experimental treatment has yielded a nonsignificant effect.

To evaluate these patterns, the observations were arranged into comparison sets of treatment (T), Hawthorne (H), and no-treatment control (C). The effect sizes associated with each comparison set were subjected to the null hypothesis test (H_0: $T - C = 0$, H_0: $H - C = 0$, and H_0: $T - H = 0$). Each of the 38 comparison sets was then placed into either the good, weak, rare or nonsignificant cases, as defined above, according to control procedure.

Frequency counts of the occurrence of each case for each Hawthorne control type are indicated in Table 5. Observations were distributed predominantly in nonsignificant cases (65.8%). A total of only three good cases across all three control procedures suggests that the position of the Hawthorne group within studies that otherwise found significant treatment effects was rarely consistent with the model.

TABLE 3
Summary of effect-size statistics (Hawthorne vs. control groups) by control procedure

Control procedure[a]	N	Unweighted		Weighted		
		M_d	S_d	M_d	S_d	95% confidence interval
Attn	10	.30	.58	.01	.08	−.13 to .15
Act	20	−.04	.31	−.02	.06	−.14 to .09
Awa	8	.18	.35	.05	.07	−.10 to .19

[a] Attn = attention, Act = activity, Awa = awareness.

TABLE 4
Summary of weighted effect-size statistics (Hawthorne vs. control groups) by moderator variables

Moderator variable	N	M_d	S_d	95% confidence interval	H_B
Activity					.45
Treatment unrelated	15	.05	.08	−.10 to .21	
Treatment related	23	−.01	.04	−.09 to .08	
Date					1.24
Pre-1975[a]	25	−.02	.04	−.11 to .07	
After 1975	13	.08	.07	−.07 to .22	
Treatment nature					.48
Setting atypical	19	−.02	.06	−.14 to .09	
Setting typical	19	.03	.05	−.07 to .13	
Age of subjects					.06
Children	27	.01	.05	−.08 to .11	
Adults	11	−.01	.06	−.13 to .12	
Expectancy					.11
Given by experimenter	17	−.01	.05	−.11 to .10	
Not given	21	.02	.06	−.09 to .13	
Experimenter controls					.44
Present	17	.03	.05	−.07 to .13	
Absent	21	−.02	.06	−.13 to .09	
Form					5.02
Journal	11	.10	.08	−.06 to .26	
ERIC	14	−.09	.06	−.21 to .02	
Dissertation	13	.06	.06	−.06 to .19	
Awareness					.63
Subjects aware	25	.02	.05	−.06 to .12	
Subjects unaware	13	−.04	.07	−.17 to .10	
Assignment of subjects					1.19
Random	22	−.03	.05	−.13 to .07	
Nonrandom[a]	16	.05	.06	−.06 to .17	

[a] The value for H_w, the test of within-cluster homogeneity, was significant, indicating that this cluster was not homogeneous.

TABLE 5
Classification of treatment /Hawthorne/ control group combinations according to control procedure

Case	Attention ($N = 10$)	Activity ($N = 20$)	Awareness ($N = 8$)	Percentage of total
Good	2	1	0	7.9%
Weak	2	3	1	15.8%
Rare	1	2	1	10.5%
Nonsignificant	5	14	6	65.8%

Activity matching yielded the greatest number of unusual cases, but the numbers were too small for any reasonable interpretation.

All of the good Hawthorne cases involved child subjects, although this represented only 12.5% of the observations involving children. Dissertations and journal articles

produced substantially more nonsignificant cases (84.6% and 72.7%, respectively) than did ERIC reports (42.9%). Further analyses of these data subdivided according to other variables did not reveal any other relevant findings.

A less rigorous approach to evaluating the pattern of effect sizes is to determine, irrespective of statistical significance, how often the effects associated with each artifact control procedure were in the direction consistent with the Hawthorne model, i.e., how often the Hawthorne effect sizes were intermediate to those of the treatment group and no-treatment controls. Examination of these data (Table 6) reveals that 60% of the attention, 30% of the activity matching, and 25% of the awareness control procedures were consistent with the model.

Discussion

In some respects the foregoing results appear discouraging. First, there was no evidence of an overall Hawthorne effect. The mean effect associated with Hawthorne manipulations was nonsignificant, and hence such groups essentially could be regarded as no different from a no-treatment control. Moreover, a detailed analysis of these studies by their control procedure, and subsequently by other moderator variables, also revealed no systematic trends to suggest a specific artifact to pose as an alternative concern.

Although one should be cautious in drawing conclusions from nonsignificant results, the conclusion we derive from an exhaustive survey of all available studies employing Hawthorne controls seems clear. There is no artifact that can be labeled the Hawthorne effect—at least not *in the manner in which it has been manipulated to date.* The concept of a Hawthorne effect originated in an attempt to summarily account for a complexly determined set of effects in a classic study. It became reified as a major artifact of concern in the behavioral sciences by a sequence of citations by prominent methodologists (Bracht & Glass, 1968; Cook, 1962; French, 1953) and authors of methods textbooks. Although the original Hawthorne experiments have been reanalyzed, possibly more often than any other social science study, few have seriously examined the empirical evidence for the methodological construct (Cook, 1967; Diamond, 1974; Schneiderman, 1977). The data presented in this study clearly indicate that there is no artifact that should be labeled *the* Hawthorne effect. There was also no evidence to support any of three distinctive subtypes of Hawthorne effects as the source of the artifact.

Comparing the three procedures employed to manipulate and control for Hawthorne effects, there was little to choose among them. Attention seemed to fare the

TABLE 6
Frequencies and percentages of the ordinal placement of the Hawthorne group by control procedure

Ordinal placement[a]	Attention ($N = 10$)		Activity ($N = 20$)		Awareness ($N = 8$)	
	f	%	*f*	%	*f*	%
T > H > C	6	60	6	30	2	25
T > C > H	1	10	9	45	2	25
H or C > T	3	30	5	25	4	50

[a] T = treatment; H = Hawthorne; C = control.

best—its mean effect size was the largest, though not significantly so; it yielded the largest number of "good" Hawthorne effects, and the majority of attention manipulations resulted in effects that were consistent with the Hawthorne model. Most often the mean effect associated with the attention control was, as expected, larger than that of the control group mean but less than that of the no-treatment control. Although attention demonstrated no potency as an artifact of general concern, a well-designed attention manipulation might have some utility as a control with the child subjects with whom the technique is most commonly used.

The activity-matching procedure was the least successful. The mean effect size associated with activity-matching manipulations was quite small and negative. The within-study analysis revealed that the largest proportion of cases were not consistent with the expected relative effects for a Hawthorne control. A subjective examination of these manipulations revealed a diverse collection of activities that had been introduced to serve as artifact controls, often without a stated rationale for their selection.

As a manipulation of artifact and as a control technique, the awareness procedure fared little better. Although the mean effect size was positive, its magnitude was small and nonsignificant. In the within-study analyses, only one of the awareness control studies resulted in an effect that was in the direction consistent with the Hawthorne control model.

Could the presence of artifact have been masked by the small number of studies available for this review? Certainly 38 studies are adequate for computations and comparisons, and they are more than are contained in a number of other meta-analyses. Because we included virtually every study that made reference to Hawthorne controls, we do not feel that there is any serious limitation on the conclusions we drew from the data.

Would meta-analysis of a larger number of studies employing similar controls not labeled Hawthorne provide evidence of artifact that eluded us in this study? For two reasons, we think not. First, it is hard to imagine that a consistent effect that we were unable to find in a set of studies bound together by a common label would emerge in a wider set of studies employing a diversity of control techniques and labels. Second, a survey of opinions of researchers employing differing labels for specially treated control groups (Adair, Sharpe, & Huynh, 1989) left an impression of multiple conceptions of what constitutes potential artifact and of how to structure control groups.

If the control procedures are ineffective, then why do they continue to be used? When Cook (1967) wrote his critical review of the Hawthorne effect, he expressed concern that many investigators inappropriately cried "Hawthorne" because it seemed convenient to blame difficulties with research results on some nebulous artifact. The present review suggests a related problem. In the studies we reviewed, many investigators seemed to have ritualistically introduced Hawthorne controls of quite different sorts in the belief that including such specially treated control groups magically protects their study against artifact. Just as there once was a false claim of Hawthorne, there has also been a false confidence in the Hawthorne control group (Payne & Brown, 1982).

The source of the problem with artifact is a lack of theory and of systematic consideration of control procedures. In contrast to the extensive discussion in method textbooks of overall design issues, and the careful attention investigators

give to their experimental manipulations, relatively little consideration is given to artifact and much less guidance is provided as to the appropriate nature and structure of a control group. Most textbooks overlook the fact that "the idea of a 'control group' is relatively complex" (Runkel & McGrath, 1972, p. 226), and their advice regarding Hawthorne procedures is as inconsistent as the practices in the studies we have reviewed (Adair, 1981).

Should educational researchers take solace in this review and now conclude that there is no artifact that they need be concerned with? Definitely not! There may well be an artifact or artifacts that the educational researcher should be concerned with; our research suggests the likelihood that none of those that have been controlled for in the past are *the* artifact source, or that the manner of previous manipulations has not been adequate to the task. We would speculate that a more likely artifact source for educational researchers is to be found in subjects' cognitions—expectations of treatment outcomes or hypothesis awareness. The former has been the artifact of concern and controversy in studies of therapy effectiveness, but there is not agreement on the appropriateness of placebo controls in such studies. Hypothesis awareness is a problem arising from laboratory experimentation that has received less attention than it deserves.

Hypothesis awareness should not be confused with mere awareness of participation in an experiment that has been associated with the Hawthorne effect. Hypothesis awareness refers to the subjects' developing an expectation or hypothesis (not necessarily the same as the experimenter's) about the purpose of the experiment as a guide to their behavior. Such an artifact has been proposed as an alternative explanation for the Hawthorne effect (Adair, 1984; Schneiderman, 1977). From this perspective, only subjects who had formulated an "expected" outcome or "appropriate" response would provide biased data. Such an approach would require a rethinking of control procedures, including the need to perfect "quasi-control" procedures (Orne, 1973) suited to this artifact.

In conclusion, the nature and sources of experiment artifact and their corresponding controls have received less attention than their importance in educational and psychological research warrants. Additional research will be needed to better understand the complexity of artifact in general and to determine whether any of the traditional control procedures have a continuing contribution to make to educational research. If the control group is "exceedingly complex" (Runkel & McGrath, 1972), then special controls for artifact are even more complex and will require much greater consideration than they have been given in the past.

Note

[1] The complete list of studies, with an indication of those that were included in the meta-analysis, may be obtained from the first author.

References

Adair, J. G. (1981, November). *The Hawthorne effect: A reinterpretation of the methodological artifact.* Paper presented at the meeting of the Society for the Social Studies of Science, Atlanta.

Adair, J. G. (1984). The Hawthorne effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology, 69*, 334–345.

Adair, J. G., Sharpe, D., & Huynh, C. L. (1989). Placebo, Hawthorne, and other artifact

controls: Researchers' opinions and practices. *Journal of Experimental Education, 57,* 341–355.

Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. *American Educational Research Journal, 5,* 437–474.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago: Rand McNally.

Campbell, E. F. (1978). The effects of learning disabilities remediation, attention training, and the Hawthorne effect on learning-disabled children. *Dissertation Abstracts International, 38,* 4423A. (University Microfilms No. DDJ-78-01406)

Chasey, W. C., Swartz, J. D., & Chasey, C. G. (1974). Effect of motor development on body image scores for institutionalized mentally retarded children. *Journal of Mental Deficiency, 78,* 440–445.

Cook, D. L. (1962). The Hawthorne effect in educational research. *Phi Delta Kappan, 44,* 116–122.

Cook, D. (1967). *The impact of the Hawthorne effect in experimental designs in educational research* (Report No. 0726). Washington, DC: U.S. Office of Education.

Diamond, S. S. (1974). *Hawthorne effects: Another look.* Unpublished manuscript, University of Illinois, Chicago Circle.

Dickson, W. J., & Roethlisberger, F.J. (1966). *Counselling in an organization: A sequel to the Hawthorne studies.* Boston: Harvard University Press.

French, J. R. P., Jr. (1953). Experiments in field settings. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences* (pp. 95–135). New York: Holt, Rinehart & Winston.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage Publications.

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. (1982). Fitting categorical models to effect sizes from a series of experiments. *Journal of Educational Statistics, 7,* 119–137.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Herman, G. N. (1982). The effect of energizing and calming movement warm-ups on third grade children's original verbal images. *Dissertation Abstracts International, 43,* 1419A. (University Microfilms No. DDJ-82-23414)

Higgins, C., & Rusch, R. R. (1965). Remedial teaching of multiplication and division: Programmed textbook versus workbook—a pilot study. *Arithmetic Teacher, 12,* 32–38.

Kalechstein, P., Kalechstein, M., & Docter, R. (1981). The effects of instruction on test-taking skills in second grade black children. *Measurement and Evaluation, 13,* 198–201.

Orne, M. T. (1973). Communication by the total experimental situation: Why is it important, how it is evaluated, and its significance for the ecological validity of findings. In P. Pliner, L. Krames, & T. Alloway (Eds.), *Communication and affect* (pp. 157–191). New York: Academic Press.

Payne, D. A., & Brown, C. L. (1982). The use and abuse of control groups in program evaluation. *Roeper Review, 5* (1), 11–14.

Pella, M. O., Stanley, J., Wedemeyer, C. A., & Wittich, W. A. (1962). The use of the White films in the teaching of physics. *Science Education, 46,* 6–21.

Roethlisberger, F. J., & Dickson, W. J. (1939). *Management and the worker.* Cambridge, MA: Harvard University Press.

Runkel, P. J., & McGrath, J. E. (1972). *Research on human behavior: A systematic guide to method.* New York: Holt, Rinehart & Winston.

Schneiderman, M. H. (1977). Hawthorne effects in remedial program for low achieving and

learning disabled children. *Dissertation Abstracts International, 37,* 4121B. (University Microfilms No. 77-4175)

Smith, M. L., Glass, G., & Miller, T. I. (1980). *The benefits of psychotherapy.* Baltimore: Johns Hopkins University Press.

Authors

JOHN G. ADAIR, Professor, Department of Psychology, University of Manitoba, Winnipeg, Man. R3T 2N2, Canada. *Specializations*: Social psychology of behavioral research, research ethics, social sciences in developing countries.

DONALD SHARPE, doctoral candidate, Department of Psychology, University of Manitoba, Winnipeg, Man. R3T 2N2, Canada. *Specializations*: Research methodology, meta-analysis.

CAM-LOI HUYNH, Assistant Professor, Department of Psychology, University of Manitoba, Winnipeg, Man. R3T 2N2, Canada. *Specializations*: Statistical methods, psychometrics.