1987

A meta-analysis of styles of supervision: A reexamination of the Hawthorne findings

Ryan Mark Cherland

Follow this and additional works at: http://scholarworks.lib.csusb.edu/etd-project

Part of the Human Resources Management Commons

Recommended Citation
http://scholarworks.lib.csusb.edu/etd-project/390

This Thesis is brought to you for free and open access by the John M. Pfau Library at CSUSB ScholarWorks. It has been accepted for inclusion in Theses Digitization Project by an authorized administrator of CSUSB ScholarWorks. For more information, please contact scholarworks@csusb.edu.
A META-ANALYSIS OF STYLES OF SUPERVISION:
A REEXAMINATION OF THE HAWTHORNE FINDINGS

A Thesis
Presented to the
Faculty of
California State University,
San Bernardino

In Partial Fulfillment
of the Requirements for the Degree
Master of Arts
in
Psychology

by
Ryan Mark Cherland
June 1987
A META-ANALYSIS OF STYLES OF SUPERVISION:
A REEXAMINATION OF THE HAWTHORNE FINDINGS

A Thesis
Presented to the
Faculty of
California State University,
San Bernardino

by
Ryan Mark Cherland
June 1987

Approved by:

Dr. Janet Kottke, Chair, Psychology
Dr. Robert Cramer, Psychology
Dr. James Rogers, Management
ABSTRACT

This thesis examined the relationship of styles of supervision based on the Hawthorne findings with productivity levels or supervisor effectiveness as the independent criterion. Meta-analytic procedures were applied to 20 studies after correcting for sampling error and measurement error. Findings suggested that there was a positive relationship between Hawthorne styles of supervision and the independent criterion. Moderator variables of the Hawthorne styles of supervision included the type of independent criterion used, the job site of the study, and whether the independent criterion was a subjective or objective measure.
ACKNOWLEDGEMENTS

I would like to thank my committee members, Dr. Janet Kottke, Dr. Robert Cramer, and Dr. James Rogers. Their help and consideration throughout this thesis was extremely helpful and I am sincerely grateful. I would also like to thank Patricia Bartell for her understanding, as well as her reading and editing of my thesis.
# TABLE OF CONTENTS

Title Page..................................................i
Signature Page............................................ii
Abstract.................................................iii
Acknowledgements.......................................iv
Table of Contents......................................v
List of Tables............................................vii
Introduction.............................................1
  The Hawthorne Experiments............................2
  Conclusions made about the Hawthorne Experiments...14
  Criticisms of the Hawthorne Studies Conclusions.....17
Reanalyses of the Original Hawthorne Data..............21
Meta-Analysis...........................................24
Different Quantitative Review Methods...................26
Criticisms of Meta-Analysis.............................31
Methods of Meta-Analysis................................36
Criteria to evaluate/conduct a Meta-Analysis..........40
Method
  Selection of Studies...................................43
  Coding of Study Characteristics.......................46
  Calculation of Effect Sizes............................47
  Cumulation of Effect Sizes...............................51
Results..................................................54
Discussion

The Hawthorne Findings ........................................... 68
The Meta-Analytic Method .......................................... 72
Summary ....................................................................... 74

Appendices

A. Earlier Criticisms of Hawthorne Research ............. 75
B. Reference List of Examined Articles ..................... 77
C. Transformation of epsilon to an effect size ........... 92
   Combining multiple effect sizes in a study ............. 92
   Cumulation of effect sizes .................................... 93
   Test for heterogeneity of variance ....................... 95
   Post hoc contrast test ......................................... 96
D. BASIC Program .................................................. 97

References ............................................................... 101
LIST OF TABLES

Table 1. Studies used in Meta-analysis.......................48
Table 2. Results of analysis of all the studies examining Hawthorne styles of supervision..................56
Table 3. Results of analyses using production as the independent criterion and those using supervisor effectiveness as the independent criterion.................................57
Table 4. Results of analyses of studies from manufacturing sites, white collar service sites, and blue collar service sites..........59
Table 5. Results of analyses of studies conducted in the 1950s and those conducted after the 1950s...............................60
Table 6. Results of analyses of studies with an average work group size less than or equal to 10 and those with an average work group size greater than 10.................................61
Table 7. Results of analyses of studies with sample sizes less than 100 and those with sample sizes greater than 100.................................63
Table 8. Results of analyses of studies with objective independent criteria and those with subjective independent criteria..................64
Table 9. Results of analyses of studies without a combined effect size and those with a combined effect size.................................66
INTRODUCTION

The first documented scientific examination of industrial relations took place at Western Electric Company's Hawthorne Works in Chicago. These first studies have proved to be of great importance to management theory and the treatment of workers, and have undoubtedly formed a valuable contribution to the science and art of human management. The research resulting from the experiments done at the Hawthorne Works became known as the Hawthorne studies. These studies concluded that output was effected by employee morale, worker solidarity, subtle social control processes, and employee attitudes and feelings (Roethlisberger and Dickson, 1939; Wardwell, 1979). These conclusions, and the evidence upon which they are based, have come under criticism on theoretical and methodological grounds (Carey, 1967; Parsons, 1974; Pitcher, 1981; Franke & Kaul, 1978; and Schlaifer, 1980).

The lack of any definitive conclusions, and the continuing controversy over the Hawthorne studies led to the questioning of many of the revolutionary conclusions developed by the original Hawthorne researchers. These conclusions changed industrial management from an idea based on the scientific management theories developed by
Taylor (Pearson, 1945) to principles of management based on human emotional and motivational factors. Korman (1971, p. 7) states, "It is from these studies that we can date the 'human relations' influence on U.S. management and some of the newer theories of effective leadership." Studies which attempted to clarify the Hawthorne research have been unable to do so. Differing theories abound as to why the production of the workers increased, with some supporting the original conclusions and others vigorously opposing the ideas of the Hawthorne researchers. One possible way to find out if the conclusions reached by the original Hawthorne researchers were accurate is to conduct a meta-analysis on those studies that have examined the effect of 'friendly supervision' or democratic leadership styles on worker output.

The Hawthorne Experiments

The Western Electric Company, in cooperation with the National Research Council of the National Academy of Sciences, planned in 1924 to examine the relation of quality and quantity of illumination to efficiency in industry (Roethlisberger and Dickson, 1939). The results of these studies were so surprising and unexpected that it was decided that further research was needed. The illumination studies and the research they inspired became known as the Hawthorne studies. A brief description and
review of each study is important for the reader's orientation of the Hawthorne studies.

**Illumination Studies**

The illumination experiments were comprised of three experiments. These experiments started in 1924 and lasted for a period of two and one-half years. The first illumination study was conducted in three departments. The first department inspected small piece parts, the second department assembled relays, and the third department wound coils. A baseline measure of production rates was taken under normal lighting conditions for each department. The illumination intensities were then increased to specified levels and new production rates were recorded. The findings showed that no clear relationship existed between production rates and illumination levels. It was decided that an additional study was needed to control or eliminate "the various additional factors which affect production output in either the same or opposing directions to that which we can ascribe to illumination" (Snow, C.E., cited in Roethlisberger and Dickson, 1939, p. 15).

The second illumination experiment was designed to prevent the inconclusiveness of the first through the use of specific control conditions. Only one department was used, one which wound small induction coils on wooden spools. The workers were divided into two groups of equal
number, equal experience, and equal average output. The test group and the control group were placed in separate buildings to prevent competition, and illumination intensities were once again increased to specific levels. The results of the study found "appreciable production increases in both groups and of almost identical magnitude. The difference in efficiency of the two groups was so small as to be less than the probable error of the value" (Snow, C.E., cited in Roethlisberger and Dickson, 1939, p. 16).

The third illumination experiment was developed to prevent any natural light from illuminating the work area as it did in the first two. The test group and control group were used as outlined in the second Illumination study. With only artificial illumination, the test group was provided with light intensity levels of ten to three foot-candles in steps decreasing one foot candle at a time. The control group was provided with a constant illumination level of ten foot candles. When the level of illumination decreased the production rates of both the test and control groups increased slowly. Only when the level reached three foot candles did the workers complain and production decrease.

An additional informal study was conducted with only two women workers who were both willing and capable operators. They were given at times, illumination as low
as 0.06 foot candles (illumination of an ordinary moonlit night). No decrease in productivity occurred even at this level. Roethlisberger and Dickson (1939) concluded:

> Although the results from these experiments on illumination fell short of the expectations of the company in the sense that they failed to answer the specific question of the relation between illumination and efficiency, nevertheless they provided a great stimulus for more research in the field of human relations. They contributed to the steadily growing realization that more knowledge concerning problems involving human factors was essential. (p. 18)

**First Relay Assembly Study**

Pennock (1930) described that light was only a minor factor in worker output and that, "It was this discovery which suggested to us the use of the experimental method in determining the various factors governing employee effectiveness (p. 298)." It was decided that a small group of workers should be used instead of entire departments to have more control over the variables effecting the workers' output, as well as the use of experimental controls, which were absent in the illumination studies (Pennock, 1930).

Roethlisberger and Dickson (1939) describe the reasoning for a small group of workers as follows:

> In a small group it would be possible to keep certain variables roughly constant; experimental conditions could be imposed with less chance of having them disrupted by departmental routines. It would also be easier to observe and record the changes which took place both without and within the individual. And lastly, in a small group there was the possibility of establishing a
feeling of mutual confidence between investigators and operators, so that the reactions of the operators would not be distorted by general mistrust. (pp. 19-20)

The researchers picked a job which was mechanized and repetitive because it was felt that industry was heading toward this type of labor. In addition it was important that the task be the same for all workers, and that the output be such that a large statistical population could be obtained for each worker. It was decided that the assembly of telephone relays, which was performed by women workers, fulfilled the requirements best. It should be noted, however, that the workers often worked on different styles of relays. To standardize output, a conversion factor was developed to make the output amounts equivalent (Roethlisberger and Dickson, 1939, pp. 26-27).

In selecting the workers it was concluded that to avoid the element of learning in the experiment only women who were thoroughly experienced in relay assembly should be chosen. Also it was felt that the workers should be willing and cooperative because the researchers wanted the workers' genuine reactions, not false spurts of production due to the experiment, nor reductions of output because of suspicion of the management's intentions. Therefore, "the method adopted for selecting such a group was to invite two experienced operators who were known to be friendly with each other to participate in the test and ask them to
choose the remaining members of the group" (Roethlisberger and Dickson, 1939, p. 21).

The group consisted of six women, five who assembled relays, and one who acted as layout supervisor (a position which consisted of minor supervision, assigning work, and obtaining parts for each assembler). In addition, there was the test room observer, a man whose duties were to keep accurate records of all that happened and to "create and maintain a friendly atmosphere in the test room" (Roethlisberger and Dickson, 1939, p. 22).

The hypotheses being examined by this first study were described by Pennock (1930) as follows: (1) Do employees actually get tired out? (2) Are rest pauses desirable? (3) Is a shorter working day desirable? (4) What is the attitude of the employees toward their work and toward the Company? (5) What is the effect of changing the type of working equipment? (6) Why does production fall off in the afternoon?

The First Relay experiment (Landsberger, 1958), as the initial study became known, consisted of thirteen periods. Period I of the experiment consisted of recording the weekly production of each worker for the two weeks before the transfer to the test room. This baseline production was used to examine the effect of any future experimental changes. Period II lasted for five weeks and consisted of
no changes except the transfer of the women to the assembly
test room. This allowed measurement of the effect of the
transfer. Period III changed the way the women were paid.
Instead of using the departmental group incentive based on
100 workers, the test workers were made into a separate
work group for the purpose of computing piecework earnings.
This allowed each worker to receive a pay which was more
closely related to her own productivity. This period
lasted eight weeks.

Period IV started the actual experimental
investigation. Two rest pauses were introduced in the work
day. The rest periods lasted for five minutes—one in
midmorning and one in midafternoon. Period V increased the
length of the rest periods from five to ten minutes.
Period VI examined the effect of six five-minute rest
periods. Midmorning and midafternoon snacks were provided
in Period VII during those break periods.

During Period VII a personnel problem with two of the
operators arose. The workers were described as having a
problem "which involved a lack of attention to work and a
preference for conversing together for considerable periods
of time" (Roethlisberger and Dickson, 1939, p. 53). It was
decided that for the best interests of the experiment the
two women should be replaced by two new workers. The
workers were transferred in the first week of Period VIII.
One of the new workers had been in the test room before as a substitute, while the other was placed there on the recommendation of her supervisors.

Period VIII (Landsberger, 1958) examined changes in the length of the working day, while keeping the rest breaks of Period VII. Instead of working until 5 p.m., the workers were allowed to quit at 4:30 p.m. Period IX ended work at 4 p.m. Period X was exactly the same as Period VII in all details. Period XI dropped the Saturday work day. Period XII's work schedule went back to that of Period I through Period III's work schedule. Period XIII then changed the work schedule back to that of Period VII and X.

What was found in all these changes was that production increased with an almost unbroken rise, period after period, in both the average hourly and total weekly output. Pennock (1930, p. 304) stated:

Now this unexpected and continual upward trend in productivity throughout the periods, even in period number 12 when the girls were put on a full 48 hour week with no rest period or lunch, led us to seek some explanation or analysis. Observation and study suggested three possible factors which might contribute to this condition:

1. Relief from cumulative muscular fatigue.
2. Change in the pay incentive.
3. Improved psychological attitude toward the work.

The first hypothesis was rejected because the output of the individual operators showed that the increase in output rates was not dependent on the day of the week nor the time
of the day. Physical examination records also showed no signs of cumulative fatigue (Pennock, 1930).

To examine the effect of the wage incentive factor, two new experiments were developed (Landsberger, 1958). These experiments, known as the Second Relay Assembly group and the Mica Splitting Test Room were ran concurrently.

**Second Relay Assembly Study**

The Second relay assembly group consisted of five experienced assemblers who were formed into a special group to be paid separately from the rest of the department. These five assemblers were placed adjacent to each other but remained in the regular department. No other changes were made. The experiment was made up of three periods: (1) a base period lasting five weeks; (2) the experimental period, which lasted nine months; and (3) a return to the old method of payment which lasted seven weeks. Production increased by thirteen percent during the experimental period.

The researchers had trouble relating the increase to only the wage incentive program, however. The output of some of the workers taken during the baseline period showed an upward trend. It was also known that there existed a rivalry with the Relay Assembly Test Room (now in its final stages) during the experiment. It was felt by the researchers that these factors could account for the
increase just as well. Another confounding factor was that the Second Relay Assembly group was regarded with jealousy by the rest of the department, which might have kept production down. (The increase of the Relay Assembly Test Room was thirty percent.) It was because of this jealousy that the experiment was discontinued to preserve morale among the other workers.

**Mica-Splitting Test Room Study**

To examine if rest pauses without change in wage incentives would effect worker output the researchers devised the Mica Splitting Test Room study (Landsberger, 1958). Five experienced workers, who were paid on an individual piecework basis, were placed in a room by themselves to split mica, which was considered one of the most desirable jobs in the Hawthorne Works plant. During this time overtime was being worked in the mica department to meet quotas. The study consisted of five periods. Period I was the baseline measure in the regular department, which lasted eight weeks. Period II involved the movement of the workers into the test room. This recording period lasted five weeks. The first experimental manipulation occurred in Period III when two ten-minute rest periods were added. This lasted twenty-nine weeks. Period IV involved no overtime work and two ten-minute rests, and lasted forty-eight weeks. Period V reduced the
work day from eight and three-fourths hours a day to eight hours a day. The work week was reduced from five and one-half days a week to five days a week, plus the two ten-minute rests. This period lasted for seventeen weeks.

The production output rose as expected until halfway through the fourth period, when it started to decline. Also, the variability of output was greater for the Mica splitting workers than for the Relay Assembly Test Room workers. The decrease in production was attributed to the worsening economic picture (Period IV occurred during June 1929 to May 1930). The greater variability in output of the workers was due to the lack of group spirit among the mica splitters, according to the researchers (Landsberger, 1958).

Additional Studies

Two additional studies conducted at Hawthorne, but only tangential to the present study, were the Interviewing research and the Bank Wiring study. The Interviewing program at Hawthorne was designed to supply case material for supervisory training courses. It was started when the Relay Assembly Test Room study was half-way completed. It covered more than 86,000 comments on 80 topics during 10,000 interviews (Landsberger, 1958). Comparisons between men and women workers showed differences in urgency rather than in tone. Men showed a greater interest in matters
which affected their and their families' economic security. Women showed a greater concern with working conditions such as overtime, fatigue, and social contact.

The Bank Wiring study examined how small working groups evolved production standards to which individuals were forced to adhere. Production records showed that most individuals had "straight-line" output curves. One week's output did not differ from other weeks' outputs. The researchers did not introduce any experimental observations except the fact that the workers were being observed. The researchers found that conformity and nonconformity to the group norms seemed to determine whether or not an operator was accepted by the group (Landsberger, 1958).

To summarize, the Hawthorne studies examined the following areas of interest (Franke and Kaul, 1978): (1) The Illumination study which examined the physical work environment. It was composed of three studies and suggested that human factors rather than physical working conditions determined worker satisfaction and performance; (2) The First Relay study which examined physical work environment, physical work requirements, management and supervision, and social relations of workers. It was the major Hawthorne experiment and concluded that increased worker performance was due to improved human relations, and to a somewhat lesser degree, rest pauses; (3) The Second
Relay study which looked at management and supervision. It suggested only a moderate effect on performance due to small group incentive pay; (4) The Mica Splitting study which examined the effect of physical work requirements and found that performance was only moderately affected by rest and shorter work periods; (5) Interviewing which studied management and supervision, and the social relations of the workers. Supported previous conclusions on the importance of social interactions to performance. Found first indications of employees restricting output due to employee interrelations; and (6) The Bank Wiring study which examined the social relations of workers and showed how employee interrelations in a large group of workers standardized the pace of work.

Conclusions made about the Hawthorne experiments

Some of the conclusions made by the Hawthorne researchers were described by Mayo (1945):

There has been a continual upward trend in output which has been independent of the changes in rest pauses. This upward trend has continued too long to be ascribed to an initial stimulus from the novelty of starting a special study.

The reduction of muscular fatigue has not been the primary factor in increasing output.

Cumulative fatigue is not present...

There has been an important increase in contentment among the girls working under test-room conditions.

There has been a decrease in absences of about 80 per cent among the girls since entering the test-room group. Test-room operators have had approximately one-third as many sick absences as
the regular department during the last six months...

Output is more directly related to the type of working day than to the number of (working) days in the week...

Observations of operators in the relay assembly test room indicate that their health is being maintained or improved and that they are working within their capacity...

The changed working conditions have resulted in creating an eagerness on the part of operators to come to work in the morning...

Important factors in the production of a better mental attitude and greater enjoyment of work have been the greater freedom, less strict supervision and the opportunity to vary from a fixed pace without reprimand from a gang boss.

The operators have no clear idea as to why they are able to produce more in the test room; but as shown in the replies to questionnaires... there is the feeling that better output is in some way related to the distinctly pleasanter, freer, and happier working conditions... (pp. 65-67)

Of these conclusions, Roethlisberger and Dickson (1939) believed that the following two warranted the highest consideration:

At least two conclusions seemed to be warranted from the test room experiments so far: (1) there was absolutely no evidence in favor of the hypothesis that the continuous increase in output in the Relay Assembly Test Room during the first 2 years could be attributed to the wage incentive factor alone; (2) the efficacy of a wage incentive was so dependent on its relation to other factors that it was impossible to consider it as a thing in itself having an independent effect on the individual. Only in connection with the interpersonal relations at work and the personal situations outside of work, to mention two important variables, could its effect on output be determined. (p. 160)
Other explanations that were developed included the fact that the workers found themselves to be experimental subjects, were under less autocratic supervision, as well as factors such as teamwork, cohesiveness, informal organization, interpersonal relationships, and social unity (Parsons, 1978). But whatever the explanation, the importance of the Hawthorne studies to management perspectives cannot be denied. As stated by Roethlisberger and Dickson (1939):

Hitherto management had tended to make many assumptions as to what would happen if a change were made in, for example, hours of work or a wage incentive. They now began to question these assumptions and saw that many of them were oversimplified. They began to see that such factors as hours of work and wage incentives were not things in themselves having an independent effect upon employee efficiency; rather, these factors were no more than parts of a total situation, and their effects could not be predicted apart from the total situation. (p. 185)

It was these two conclusions by Roethlisberger and Dickson which developed into the fundamental writings of industrial and social relations. Books such as *Management and the Worker* (Roethlisberger and Dickson, 1939), *Management and Morale* (Roethlisberger, 1941), *The Industrial Worker* (Whitehead, 1938), *Leadership in a Free Society* (Whitehead, 1936), *The Human Problems of an Industrial Civilization* (Mayo, 1933), and *The Social Problems of an Industrial Civilization* (Mayo, 1945)
describe the research conducted at Hawthorne and its implications for management. The influence of these writings on Industrial/Organizational psychology and management theory changed the emphasis of management from one based on simple pay incentives to one based on human relations, styles of leadership, group standards, and other social factors on work performance—to the exclusion of almost all other approaches throughout the 1950s and 1960s.

**Criticisms of the Hawthorne Studies' Conclusions**

Although the conclusions of the Hawthorne studies did not escape complete critical analysis (Landsberger, 1958), it was not until the mid 1960's when the Hawthorne studies were brought back under scrutiny. Some of the more recent articles which have examined the Hawthorne studies and their conclusions will be reviewed. Since these articles review previous research, it is felt that the older studies would be redundant in many cases. This is by no means an exhaustive review, but is intended to provide the reader with the most recent analyses which are representative of the field. (For a list of the earlier criticisms, the reader is referred to Appendix A.)

The first relatively recent criticism of the Hawthorne studies was conducted by Carey (1967) in which he concluded that the decisions reached by the Hawthorne investigators were not supported by the evidence. Carey felt that it was
the material rewards of money which influenced work morale and behavior the most. Carey supported his conclusion with the following factors: (1) Apart from a slight increase in output (4-5%) following the introduction of the preferred incentive system, there was no increase in weekly output during the first nine months of the study. (2) When it became apparent that free and friendly supervision was not getting results, discipline was tightened, which led to the dismissal of two of the five women. (3) The dismissed workers were replaced by two women of special motivation and character who immediately led the rest of the group in a sustained increase in output. One of the two women who had a special need for money rapidly undertook a strong disciplinary role in regard to the rest of the group. (4) Output only showed an increase when the two women with the lowest output were replaced with the two, new motivated workers who accounted for the major part of the groups' increase. (5) Only after the two new women arrived and the resulting increase in output occurred, did supervision once again turn friendly and relaxed. There is no evidence that output increased because supervision turned friendly.

Carey (1967, p. 416) concludes that "far from supporting the various components of the 'human relations approach,' [the results] are surprisingly consistent with a rather old-world view about the value of monetary
incentives, driving leadership, and discipline."

Carey's criticism is minimized by Shepard's (1971) response in which Shepard contends that the Hawthorne researchers did not try to reduce the effects of financial rewards and overemphasize the effects of friendly supervision. In fact, Shepard shows how Mayo acknowledges that the results could not be due simply to differences in supervision, but resulted from something more, which Mayo described as 'human situations.' Shepard concluded that a primary contribution of the Hawthorne researchers remains their attempt to place financial incentives in a social context.


...these two pioneers were probably important in preserving a view of workers as irrational and unintelligent and of the capitalist factory as nonexploitative and free of class conflict. This view, which is clearly identified with defense of the capitalist mode of production, persists to the present time in discussions of the psychology of industry and particularly in reference to the Hawthorne research. (p. 867)

Bramel and Friend support their viewpoint with a review of instances of worker resistance at Hawthorne and the way the experimenters handled the resistance (e.g., the replacement of two workers during Period VIII). Bramel and
Friend also provide evidence that Mayo and Roethlisberger trivialized and hid any worker discontent through their explanations of the workers' statements and actions as being irrational, emotional, and based on misunderstandings between management and the workers. Bramel and Friend believe that Mayo and Roethlisberger have misstated the facts to provide validity to the capitalistic social order over a Marxist explanation that class bias existed at Hawthorne.

Bramel and Friend's analysis has been criticized by many researchers (Toch, 1982; Stagner, 1982; Parsons, 1982; Feldman, 1982; Locke, 1982; Vogel, 1982; and Sonnenfeld, 1982). All disagree with the Marxist content used by Bramel and Friend as being inappropriate for an explanation of what occurred at the Hawthorne factory, and some state that Bramel and Friend are also guilty of misstating the facts to support their conclusions. Given Bramel and Friend's ideological analysis and the extensive criticisms, there are two valid points which Bramel and Friend have made: (1) That there were management pressures to increase production on the experimental workers during the experiment, and (2) Mayo and Roethlisberger glossed over problems with worker discontent during the experiments, which could have influenced experimental results.

Another article (Parsons, 1974) which questioned the
conclusions of the Hawthorne researchers decided that the Hawthorne effect was in actuality operant conditioning. This conclusion is based on the fact that the workers were notified of their work output. Given this occurrence, Parsons believed that a combination of information feedback and financial reward caused operant conditioning to occur which was seen in the progressive increases in response rate.

Balling, Weiss, and Steigleder (1985) explain the increase in production at Hawthorne using a NeoHullian learning theory model, which states that behavior can be motivated by aversive drives such as fear, frustration, altruistic drive, and effectance. The experimental workers were placed in conditions which Balling et al. viewed as being aversive (i.e., constant supervision, constant evaluation, managerial discipline, fear of being laid-off, etc.). These aversive conditions motivated the workers to work at higher production levels, and therefore, account for the "Hawthorne effect." Although this view is similar to that of Mayo and Roethlisberger in that it emphasizes attention, Balling et al. believe that the attention was an aversive motivator while Mayo and Roethlisberger viewed the attention as being a positive motivator.

**Reanalyses of the Original Hawthorne Data**

The first statistical analysis of the Hawthorne studies
was conducted by Franke and Kaul (1978). The original data of the First Relay experiment was statistically analyzed using time-series econometric techniques. These techniques allow specific examination of even inadvertent experimental changes. Also, through the use of serial correlation, the influence of historical factors could be measured. Franke and Kaul at first performed zero-order correlations for the first relay group—using all the periods, then separating those seven periods prior to the replacement of the two unsatisfactory workers, and then the periods after the replacement. Next, the best multiple regressions were determined using the three production measures (hourly output, weekly output, and repair time), with correction for serial correlation where necessary. Finally, Franke and Kaul, forced the group models on the individual workers' data and then determined whether alternate models provided a better explanation of the variance. It was found that experimental variables accounted for over 90% of the variance in quantity and quality of output. The variables of managerial discipline, economic adversity, time set aside for rest, and quality of raw materials explained most of the experimental variance.

Schlaifer (1980) provided an alternative statistical analysis which used models in which the average productivity increased gradually over time instead of in
abrupt jumps as productivity did in the model developed by Franke and Kaul (1978). Schlaifer's models found discipline to be less of a factor than Franke and Kaul's model. This was attributed to Franke and Kaul's model not including a smooth function of time among the predictor variables for their stepwise regression program. Because this predictor variable was not included, their program used the discipline variable as a proxy for the smooth function of a time variable. The relationship between discipline and the smooth trend variable was found to be a correlation coefficient of .84 (Schlaifer, 1980).

Pitcher (1981) reported statistical evidence for a learning model. By fitting a standard learning equation to the data for each operator, evidence was found to support the conclusion that the increased output at Hawthorne was due to conditions which motivated increased learning. These conditions were due to better reinforcement situations such as increased status and new economic incentives, and regular performance feedback.

All these alternative explanations for what happened at the Hawthorne Works point away from the original conclusions made by the Hawthorne researchers, that friendly supervision had a significant impact on the output of workers. Because the impact of the conclusions made at Hawthorne are so important to management theory, and given
the many conflicting reanalyzes of the Hawthorne research, it is desirable to examine whether or not additional research on the effect of friendly supervision supports or refutes the original Hawthorne conclusions. To do such an examination one must use the studies which have already been conducted and combine them in some understandable manner. Recently there has developed a quantitative method in which a large amount of previous research can be analyzed as a single database. This quantitative review method has become known as meta-analysis.

Meta-Analysis

The term meta-analysis was first used by Glass (1976) to describe his method of a quantitative review of research literature. This type of review is also called research integration, and quantitative assessment of research domains (Green and Hall, 1984). Meta-analysis is nothing more than one out of several qualitative methods in which research findings are summarized and reviewed. The difference is that meta-analysis uses numbers and statistical methods for organizing large quantities of data and then extracts information from the data. It allows researchers to generalize without doing "violence to a more useful contingent or interactive conclusion" (Glass, McGraw, and Smith, 1981, p. 55). Instead of adding to a nearly incomprehensible mass of previous research, one can
examine a research area scientifically, discover what has occurred significantly in the past, locate any connections between the experiment effect and methodological conditions, and then generalize to the appropriate domain and suggest areas of future research. Meta-analysis was developed to provide reviewers with an objective method of review in comparison to traditional methods.

The literature supports the idea of a quantitative review method. Sawyer (1966) found that the statistical modes of both data collection and combination were superior to the clinical methods which relied on subjective judgements. A more comprehensive analysis was performed by Jackson and his conclusions were cited by Glass et al. (1981) as follows:

(a) Reviewers frequently fail to examine critically the evidence, methods and conclusions of previous reviews on the same or similar topics. [Although 75 percent of the reviewers cited previous reviews, only 6 percent examined them critically.]

(b) Reviewers, often focus their discussion and analysis on only a part of the full set of studies they find, and the subset examined is seldom a representative sample and it is seldom clear how it (the subset) was chosen. [Only 3 percent of the reviewers appeared to have used existing indexes—e.g., ERIC—in their search; only 22 percent selected a fair sample of studies, in the judgement of Jackson's coders; and only 3 percent analyzed the full set of studies found.]

(c) Reviewers frequently use crude and misleading representations of the findings of the studies. [About 15 percent of the reviewers classified studies according to whether their findings were 'statistically significant'; ...frequently,
reviewers report test statistics ($t$, $F$, etc.) for one or more studies.

(d) Reviewers sometimes fail to recognize that random sampling error can play a part in creating variable findings among studies.

(e) Reviewers frequently fail systematically to assess possible relationships between the characteristics of the studies and the study findings. [Fewer that 10 percent of the reviewers studied whether the findings of the research were mediated by characteristics of the persons studied, the study context, the nature of the experimental intervention or the characteristics of the research design.] The lack of systematic examination of these relationships is important because reviewers frequently eliminate studies from consideration because of a priori judgements that their findings are flawed by one or another study characteristics.

(f) Reviewers usually report so little about their methods of reviewing that the reader cannot judge the validity of the conclusions. (p. 13)

**Different Quantitative Review Methods**

One of the first quantitative review methods developed was vote-counting. This method required the reviewer to classify the literature as supporting, conflicting, or neither supporting nor conflicting. The category which contained the most "votes" or number of studies was taken as the explanation for the phenomenon (Light and Smith, 1971). This technique did not take into account either the experiments' sample sizes nor experimental effect sizes and this reduced its effectiveness. When one does not take effect size into account, the degree to which the examined phenomenon is present in the population of interest is not known nor understood.
The importance of effect size should not be ignored. Cohen (1977, pp. 9-10) described effect size as "the degree to which the phenomenon is present in the population or the degree to which the null hypothesis is false." The larger the effect size the greater is the manifestation of the experimental effect, and intuitively, the greater the experiment's power. In other words, if there are differences in the experimental population, they will be found. Cohen (1979) states:

The larger the ES[effect size] posited, other things (significance criterion, sample size) being equal, the greater the power of the test. Similarly, the relationship between ES and necessary sample size: the larger the ES posited, other things (significance criterion, desired power) being equal, the smaller the sample size necessary to detect it. (p. 11)

Effect size is calculated using different equations for different statistical tests (t, Fs, rs, etc.). Glass et al. (1981) state that the most informative and straightforward measure is the mean difference between experimental and control groups divided by within-group standard deviation:

$$d = (\bar{x}_E - \bar{x}_C)/s_x$$

Whatever the method used, the need for studies to publish effect sizes is emphasized in one instance by Rosenthal (1978, p. 192), "We owe it to our readers to give...an estimate of the probable size of the effect in terms of a sigma unit, a correlation coefficient, or some other
estimate. Providing the readers effect size measurements allows the readers to make judgements about the experimental effect and the relationship of the dependent and independent variable(s). It also allows one experiment to be compared with similar experiments because the effect size can be transformed into t and r statistics, which allows it to be compared across different statistical tests.

Getting back to quantitative reviews, another method involves the cumulation of significance levels across studies. The probabilities obtained from two to more studies are combined by using methods which included adding logs, adding probabilities, adding t's, adding Z's, adding weighted Z's, testing the mean p, testing the mean Z, counting, and blocking (Rosenthal, 1978). But again, these methods do not include any way of measuring effect sizes which hinders their use.

Meta-analysis is viewed as the most sophisticated literature review method available (Tannenbaum and Jones, 1983). Through the use of a systematic, comprehensive review of all related literature, both published and unpublished, the reviewers collect and convert the effect sizes. The conversion of the effect sizes to a common metric, such as a correlation coefficient, r, allows for different experiments, using different statistical
measures, to be compared. The effect sizes are coded, that is, recorded onto a coding sheet which lists other variables of interest as well. The coding sheets are devised by the person conducting the meta-analysis and classify the characteristics of the study into two broad areas, substantive and methodological. Substantive variables of interest are those characteristics of the study that are specific to the problem being studied. In a study which examines a drug treatment for example, substantive characteristics would include the type of drug, the size of the dose, and the age of the subject, to name a few. Methodological variables of interest are those characteristics of a study which deal with the research methods of the studies (Glass et al., 1981). Examples of methodological characteristics include sample size, test reliability, randomization versus matched versus nonequivalent groups, etc. This coding is standardized and applied to each study. When the studies have been coded the effect sizes are regressed against the variables of interest. This allows the variables to determine the factors which contribute to the variance of results across studies. Methods other than regression may be used but as with any statistical technique, the methods should be fully supported by the literature (Tannenbaum and Jones, 1983).

An example of a meta-analysis is given by Glass et al.
(1981, pp. 26-31). Twelve studies were found to have tested the effects of psychotherapy on asthma. Eleven of the studies involved the use of treatment and control groups, two others used pretest-posttest designs. The effect size was calculated by subtracting the control mean from the treatment mean and dividing the difference by the control group standard deviation. The results of the meta-analysis showed that the mean standard deviation was 0.86, which means that the average subject who received psychotherapy was 0.86 standard deviations above the mean of the control groups.

The relationship between the effects of psychotherapy and some of the features of the therapy and the patients were examined. It was found that there was no significant difference between the type of psychotherapy and the effect sizes. There was a significant effect found in the age of the patient and the mean effect size, with a linear correlation of .40 being found. Hours of therapy was found not to be significant nor was the number of weeks of post therapy. This study showed that psychotherapy, primarily behavioral therapies and hypnotherapy, have a large ameliorating effect on asthma sufferers, with the age of the patient contributing to the effectiveness of the treatment.

Meta-analysis is therefore an effective means of
combining a large number of studies into a single data-base which allows a reviewer to perform an indepth analysis on research in a specific area of interest. There are however, many who feel that there are no valid reasons to support the aggregation of studies into a single database. In addition, the researcher conducting a meta-analysis is using studies which possibly define variables differently. The arguments against meta-analysis should therefore be examined.

**Criticisms of Meta-Analysis**

Arguments against meta-analysis were given by Glass et al. (1981, p. 218) as:

1. **The Apples and Oranges Problem.** It is illogical to compare 'different' studies, that is, studies done with different measuring techniques, different types of persons, and the like.
2. **Use of Data From 'Poor' Studies.** Meta-analysis advocates low standards of quality for research. It accepts uncritically the findings from studies that are poorly designed or are otherwise of low quality. Aggregated conclusions should only be based on the findings of 'good' studies.
3. **Selection Bias in Reported Research.** Meta-analysis is dependent on the finding that researchers report. Its findings will be biased if, as is surely true, there are systematic differences among the results of research that appear in journals versus books versus theses versus unpublished papers.
4. **Lumpy (Nonindependent) Data.** Meta-analyses are conducted on large data sets in which multiple results are derived from the same study; this renders the data nonindependent and gives one a mistaken impression of the reliability of the results.
These criticisms are countered by the following arguments: (1) Glass et al. (1981) argue that the studies compared in a meta-analysis are similar to the way in which different subjects are compared in traditional research. That is, a researcher who criticizes the pooling of the results of one, five, or ten, studies should explain why there is nothing objectionable in the pooling of the results from one, five, ten, or one hundred subjects. Another counter to this criticism is there is no need to compare studies that are the same because the results should be the same as well, within statistical error. It must be remembered that the formulation of a review is contingent on the nature and scope of the question or hypothesis being examined. If the hypothesis being examined requires a global analysis, one can always later stratify the sample into smaller, more homogenous groups for a more conceptually refined analysis.

(2) Although meta-analysis uses studies of "poor" quality, it does not advocate poor research design. The methodological strength of each study can be coded and taken into consideration in the analysis. However, Hunter and his colleagues (Hunter et al., 1982) have developed a method which has the added advantage of recognizing and correcting for some of the artifactual and methodological problems affecting the results of the studies to be
A Schmidt and Hunter type of meta-analysis is based on the idea that much of the variation in results across samples or studies is due to statistical artifacts and methodological problems rather than to truly substantive differences in underlying population correlations. There are three types of error variance which can be corrected by the literature reviewer: (a) sampling error due to differences in sampling size, (b) measurement error due to imperfect instruments, and (c) range variation, which occurs when the independent variable varies more or less in the population being studied than in the reference population. One other type of error is reporting error. This includes incorrect computations, typographical errors, and the like, but this type of error is uncorrectable without examining the original data for each study.

(3) Although there do exist differences in what medium a study is published in and its research results, this cannot be considered a cogent criticism of meta-analysis, which was able to demonstrate the existence of such biases. Selective publication can be dealt with using meta-analytic procedures by collecting all of the research and analyzing it separately by mode of publication. In addition, when published data sets are compared to large unpublished data sets (U.S. Government studies), they are very similar in
terms of means and variance (Schmidt, Hunter, Pearlman, and Hirsh, 1985). Rosenthal's (1979) method to determine the number of unlocated studies averaging a mean $Z$ of .00 that would be required to change the conclusions of a quantitative research review show that the required number of missing studies are usually so many as to have little possibility of existing (e.g., 200 to 10,000+).

Unpublished studies are not usually the result of nonsignificant results but rather methodological weaknesses (Schmidt and Hunter, 1977). Hunter et al. (1982, p. 30) state, "It seems likely that most of the difference between the average effect size of the published and unpublished studies is due to differences in the methodological quality. If attenuation effects were properly corrected for, differences might disappear."

(4) When multiple results are derived from the same study, this causes the data to become nonindependent. This criticism is acknowledged as being quite cogent. Nonindependence reduces the reliability of estimation of averages or of regression equations. One solution to the problem would be to average all findings within a study up to the level of the study and proceed with a meta-analysis with "studies" as the unit of analysis, instead of the number of effects. In some cases, however, this technique will hide many important questions from analysis. An
alternative solution is to utilize analytic procedures that take the problem of nonindependence into account, but the cost is a more complex analysis and greater conceptual distance from the original constructs of interest. A further alternative is to specify *a priori* a particular measurement instrument or type of instrument and to select only that type when multiple measures are used. For example, one would select a particular measure which is common in a particular domain of research or a particular type of measure common to the research area. Only data from these measures would be used even if several other types were reported in the study.

The most conservative approach is to record the effect size for each measure and then pool the estimates of multiple effect sizes in each study in the initial analysis (Rosenthal, 1984). A recent method to combine the multiple effect sizes of a study has been developed by Rosenthal and Rubin (1986). This method incorporates the degrees of freedom and the intercorrelation among the dependent variables, which provides a more accurate and useful summary effect size.

Tannenbaum and Jones (1983, p. 13) feel that "none of these reasons seem severe enough to prevent meta-analysis from becoming a more prevalent research tool." By examining experiments in an objective, scientific manner,
the researcher can make statements supported by the aggregation of data. This is to be preferred over those statements which describe previous research in general terms but decide that specific and definitive conclusions from a general survey of the literature can be made. Meta-analysis not only offers a way to cumulate the results of many studies but it also allows an examination of those methodological factors of experiments which effect the results, something that no other methods of literature reviews offer.

**Methods of Meta-Analysis**

Although all methods of meta-analysis share some common attributes, such as standardization, comprehensiveness, documentation, and quantification, there are differences in the two main approaches, one which was developed by Glass, and the other by Hunter.

Glass, who first used the term meta-analysis to describe his quantitative technique for research review, emphasizes a need for the computation of effect sizes when integrating studies. This allows the researcher to examine how study characteristics affect summary findings. It also provides an estimate of overall mean and variance of the effect sizes (Glass et al., 1981). Glass' approach encourages a complete search through existing sources for studies. Coding is then conducted on all factors which might
influence the experimental effect. Interrater reliability is calculated for the coders to insure uniformity in the procedure. The experimental effect sizes for each study is then calculated into a common metric. Methods for these conversions are listed by both Glass and Hunter (Glass et al., 1981; Hunter, Schmidt, and Jackson, 1982). Briefly, however, some of the conversions are listed below:

\[ r = \sqrt{\frac{t^2}{t^2 + df \text{ error}}} \]

\[ r = \sqrt{\frac{F(1,df \text{ error})}{F(1,df \text{ error}) + df \text{ error}}} \]

\[ r = \sqrt{\frac{z^2}{N}} \]

Once the database is established, any appropriate statistical procedure may be used to assess which factors contribute to the variance of results across studies.

Schmidt and Hunter developed their meta-analytic techniques as an extension of their work in the area of validity generalization procedures for employment tests. Hunter et al. (1982) follow the same procedures as Glass except for the inclusion of procedures to correct for artifactual variance before the coding of the moderator
variables. These potential sources of artifactual variance include: sampling error, differences across studies in the reliabilities of predictors and criterion measures, and the differences between studies in the degree of range restriction. Other sources of artifactual variance exist, but estimates of these sources are considered unobtainable. The procedures are done to remove as much error variance as possible for the observed findings. When the researcher has corrected for sources of artifactual variance, he/she then selects those moderator variables which have logical, statistical, or psychometric justifications with which to be included in the analysis. Hunter et al. have criticized the Glass' approach of coding from 50 to 100 study characteristics as capitalizing on chance factors due to the large number of characteristics that were coded. However, Hunter et al. do concede that if the estimated variance of effect sizes across studies is substantially greater that zero after corrections for artifacts have been made, then Glass' approach would be a supplemental step to the Hunter et al. procedures.

These two meta-analysis procedures approach literature reviews from different philosophical perspectives according to Mathieu and Tannenbaum (1983):

One might regard the Hunter approach as a confirmatory procedure aimed at specifying "what is known", while the Glassian method seems closer to exploratory techniques which permit unexpected
relationships to surface (p. 7).

These differences in approach should not limit the meta-analyst to restricting the quantitative review to only one way of analysis, and Mathieu and Tannenbaum (1983) recommend a combination of the two approaches which allows for an a priori specification of the primary variables as well as the listing of secondary or exploratory variables. A meta-analysis done in this way will result in confirmation or denial of existing data, as well as point to areas of future interest.

Statistical models have been developed for effect size analysis (Strube, Gardner, & Hartmann, 1985). A meta-analysis on effect sizes infers that the data are derived from the population of studies which can be accurately described by a statistical model. It is assumed that the meta-analyst wishes to treat the studies under review as a sample of observations concerning the true effectiveness of a treatment. The evidence provided by the sample is then used to estimate the true values of parameters characteristic of the population. Hedges (1981, 1982a, 1982b, 1982c, 1983) has distinguished two general types of effect size, a fixed effects model and a random effects model. The difference between the two is analogous to that made in the analysis of variance. Strube et al. (1985) have summarized the models in the following manner:
In the fixed effects model, the studies can be viewed as random samples from a population characterized by a single, fixed effect size. Under the assumption of valid measurement without error (a rather untenable assumption in practice) sample effect sizes will deviate from this fixed population value as a function of sampling error. Furthermore, it is possible for more than one fixed population value to exist and be estimated by the sample. In this latter instance, variability in sample estimates reflects an additional component (i.e., that there are two or more parent populations). For example, diverse treatments may be represented by discrete population effect sizes. An alternative random effects model proposes that the population effect size is randomly distributed with its own mean and variance, rather than having a fixed value. Thus variability in sample effect sizes reflects not only sampling error, but variability in the parent population as well (i.e., the sample values are estimating different population values) (pp. 71-72).

**Criteria to evaluate/conduct a Meta-Analysis**

To perform an evaluation of the literature as involved as a meta-analysis, one must follow some guidelines and suggestions which prevent common mistakes. In an attempt to replicate a meta-analysis, Bullock and Svyantek (1985), identified some potential problems they encountered. These problems included the public availability of a list of the studies used, lack of adequate coding documentation and decision rules, a narrow domain of generalization (i.e., a meta-analysis was done which used only studies found in one specific journal), inadequate coding of study characteristics, selective reporting of results, and the overinterpretation of the results.
Therefore, one should be able to conduct a meta-analysis which is informative and helps describe a research area by holding the coding of the studies' results to a minimum to prevent overinterpretation and capitalizing on chance factors, as well as correcting for sampling error and measurement error. By holding to these criteria, it is felt that a meta-analysis on studies which have researched the relationship between "friendly supervision" and production or some other measure of supervisor effectiveness will help in clarifying the conclusions of the Hawthorne researchers regarding friendly supervision.

The basis for much of management theory involves the "friendly supervision" hypothesis put forth by the Hawthorne researchers, and deserves a thorough review which can only be provided by a quantitative literature review.

This meta-analysis will only examine studies which were conducted on supervisors in the work place. This group of studies has settings which are the most similar to the original Hawthorne environment. Since the purpose of the study is to find either support for or to dispute the original Hawthorne conclusions, this study will not examine any research which was in a laboratory setting or which examined leadership styles as opposed to supervisory styles. Although this limits the size of the database it is felt that given the purpose of this study the
limitations are appropriate.
METHOD

Selection of Studies

An attempt was made to locate, summarize, and analyze the results of all published studies reporting the effects of different styles of supervision — in particular, those judged to be based on human relations methods — on production levels or some measure of supervisor effectiveness. Some of the possible characteristics of a human relations manager are discussed by Gordon (1958):

(1) He permits all members to discuss policy formation. Encourages the group to make necessary decisions.
(2) He permits discussions of future as well as present activity. Does not try to keep members in the dark about future plans.
(3) He permits members to define their own job situation as much as possible. For example, the defining of the way to accomplish tasks and the division of tasks is left up to the group.
(4) He focuses on obtaining objective facts on human problems. Tries to base any necessary praise or discipline upon these objective facts and not upon his personal needs. (p. 420)

A limitation on the sample of studies to be selected was that they were conducted in an actual work setting and not be a laboratory experiment. This limitation was used because the Hawthorne findings were based in a work setting and this meta-analysis is an attempt to examine the strength of the Hawthorne findings using results from
similar environments. Another requirement was that the studies had to have usable statistical analyses to calculate an effect size. A usable statistic was considered any measure which describes the relationship between any variable X and any variable Y. Just about any test statistic can be converted into a usable effect size measure. Even studies which list only a significance level can have an effect size approximated given the size of the sample. Most studies which were not included in the final sample were discarded because of their failure to meet the setting or general topic requirement rather than the statistical requirement.

A literature search was performed manually on Psychological Abstracts through the years 1945 to 1969. It was felt that since the Hawthorne findings were not made public until 1939 (Management and the Worker, Roethlisberger and Dickson, 1939) and with the advent of World War II, few, if any, studies would be conducted and published before 1945 (it was found that the earliest usable study was published in 1952, giving some credence to this sampling criterion). The manual literature search ended with 1969 journal articles because a computer literature search, which was conducted, would cover the years after 1969.

The computer literature search used the DIALOG
Information Retrieval Service on the PSYCINFO, ABI/INFORM, MANAGEMENT CONTENTS, and the DISSERTATION ABSTRACTS ONLINE databases. PSYCINFO covers the world's literature in psychology and related disciplines in the behavioral sciences from 1967 to the present, and scans over 900 periodicals and 1500 books, technical reports, reports, and monographs each year. ABI/INFORM covers the literature from 1971 to the present and examines more than 500 publications in business and management. MANAGEMENT CONTENTS covers the years 1974 to the present and specializes in current information from approximately 500 journals, proceedings, and transactions involving the areas of accounting, finance, industrial relations, and organizational behavior. DISSERTATION ABSTRACTS ONLINE covers virtually every American dissertation accepted at accredited institutions since 1861.

A total of 604 abstract entries dealing with supervision were found through the manual literature search. 139 of these were found to deal with an applied work setting and evaluate supervisory styles in some manner. These were judged to be studies of possible use in the meta-analysis. The computer literature search produced an additional 70 listings of which 40 were judged to be worthy of further examination. These 179 studies are listed in Appendix B.
An effort was made to locate the desired studies to examine whether or not they could be included in the meta-analysis. Twenty-four studies were dropped because they were dissertations and it was deemed financially unfeasible to try to obtain them. Upon further examination of the remaining studies it was found that only 20 articles were set in a work setting, examined human relations styles' of management, and contained the necessary statistics for calculating an effect size.

Coding of Study Characteristics

The collected articles were coded for certain characteristics which were believed to have possible moderating effects on the findings. These characteristics can be classified into two broad areas, substantive and methodological. The substantive characteristics which were coded for were: (1) whether the independent criterion was productivity of the work group or effectiveness of the supervisor; (2) whether the study was situated in a manufacturing, blue collar-service, or white collar-service setting; (3) whether the study was published in the 1950s or whether it was published after the 1950s; and (4) whether the average work group in the study had more than ten workers or whether the work group had ten or less workers. The substantive characteristic 1950s vs post-1950s was included because the human relations school
of thought became popular during the 1950s and it was thought that there might be a potential backlash away from human relations management styles after the 1950s by researchers. The substantive characteristic dealing with the work group size (also known as "tallness" of an organization) was examined to see if the size of a work group could affect the effectiveness of a "friendly style" of supervision, with the smaller work groups being more positively influenced than the larger work groups.

The methodological characteristics examined were: (1) whether the sample size was under 100 or whether it was over 100; (2) whether the independent criterion was a subjective measure or an objective measure; and (3) whether there were multiple effect sizes in the study or whether there was only one effect size per study. A listing of the sample studies with their coded characteristics appears in Table 1.

**Calculation of Effect Sizes**

An effect size was calculated for each study based on $r$. The majority of the studies used $r$ which was why this statistic and not $d$, which is the statistic preferred by Glass, was used. Some of the studies conducted in the 1950s used $\varepsilon$ as their statistical measure. Since this statistic is relatively rare today, a transformation to an $r$ value was not listed by any of the sources of
<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Sample Size</th>
<th>Effect Size</th>
<th>Type of Criterion</th>
<th>Job Site</th>
<th>Average Group Size</th>
<th>Objective Measure</th>
<th>Used Effect Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patchen (1962)</td>
<td>700</td>
<td>.33</td>
<td>P</td>
<td>M</td>
<td>&gt;</td>
<td>S</td>
<td>W/O</td>
</tr>
<tr>
<td>Indik, Georgopoulos, &amp; Seashore (1961)</td>
<td>975</td>
<td>.15</td>
<td>P</td>
<td>BC-S</td>
<td>&gt;</td>
<td>0</td>
<td>W/</td>
</tr>
<tr>
<td>Comrey, Pfiffner, &amp; Beem (1952)</td>
<td>34</td>
<td>.60</td>
<td>SE</td>
<td>WC-S</td>
<td>≤</td>
<td>S</td>
<td>W/</td>
</tr>
<tr>
<td>Carp, Vitola, &amp; McLanathan (1963)</td>
<td>41</td>
<td>.65</td>
<td>P</td>
<td>WC-S</td>
<td>--</td>
<td>0</td>
<td>W/O</td>
</tr>
<tr>
<td>Nagle (1954)</td>
<td>208</td>
<td>.82</td>
<td>P</td>
<td>WC-S</td>
<td>&gt;</td>
<td>S</td>
<td>W/O</td>
</tr>
<tr>
<td>Gekoski (1952)</td>
<td>200</td>
<td>.07</td>
<td>P</td>
<td>WC-S</td>
<td>≤</td>
<td>S</td>
<td>W/O</td>
</tr>
<tr>
<td>Wilson, High, Beem, &amp; Comrey (1954)</td>
<td>163</td>
<td>.30</td>
<td>P</td>
<td>BC-S</td>
<td>≤</td>
<td>0</td>
<td>W/</td>
</tr>
<tr>
<td>Argyle, Gardner, &amp; Cioffi (1957)</td>
<td>90</td>
<td>.18</td>
<td>SE</td>
<td>M</td>
<td>--</td>
<td>S</td>
<td>W/</td>
</tr>
<tr>
<td>Dunteman &amp; Bass (1963)</td>
<td>27</td>
<td>.42</td>
<td>SE</td>
<td>WC-S</td>
<td>--</td>
<td>S</td>
<td>W/O</td>
</tr>
<tr>
<td>Bass (1958)</td>
<td>42</td>
<td>.32</td>
<td>SE</td>
<td>M</td>
<td>--</td>
<td>S</td>
<td>W/O</td>
</tr>
<tr>
<td>Author(s)</td>
<td>Sample Size</td>
<td>Effect Size</td>
<td>Type of Independent Criterion</td>
<td>Job Site</td>
<td>Average Group Size</td>
<td>Objective/Subjective Measure</td>
<td>Used Combined Effect Sizes</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------</td>
<td>-------------</td>
<td>-------------------------------</td>
<td>----------</td>
<td>-------------------</td>
<td>-----------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>Comrey, Pfiffner, &amp; High (1954)</td>
<td>93</td>
<td>.19</td>
<td>SE</td>
<td>M</td>
<td>&lt;</td>
<td>S</td>
<td>W/</td>
</tr>
<tr>
<td>Parker (1963)</td>
<td>1716</td>
<td>.13</td>
<td>P</td>
<td>BC-S</td>
<td>&gt;</td>
<td>0</td>
<td>W/O</td>
</tr>
<tr>
<td>Comrey, High, &amp; Wilson (1955)</td>
<td>29</td>
<td>.37</td>
<td>P</td>
<td>M</td>
<td>--</td>
<td>0</td>
<td>W/</td>
</tr>
<tr>
<td>Tjosvold, Andrews, &amp; Jones (1983)</td>
<td>310</td>
<td>.56</td>
<td>P</td>
<td>WC-S</td>
<td>--</td>
<td>S</td>
<td>W/O</td>
</tr>
<tr>
<td>Russel &amp; Farrar (1978)</td>
<td>507</td>
<td>.51</td>
<td>SE</td>
<td>WC-S</td>
<td>--</td>
<td>S</td>
<td>W/</td>
</tr>
<tr>
<td>Student (1968)</td>
<td>486</td>
<td>.32</td>
<td>P</td>
<td>M</td>
<td>&gt;</td>
<td>0</td>
<td>W/O</td>
</tr>
<tr>
<td>Taylor, Parker, Martens, &amp; Ford (1959)</td>
<td>65</td>
<td>.13</td>
<td>SE</td>
<td>WC-S</td>
<td>--</td>
<td>S</td>
<td>W/O</td>
</tr>
<tr>
<td>Wilson, Beem, &amp; Comrey (1953)</td>
<td>1022</td>
<td>.33</td>
<td>P</td>
<td>BC-S</td>
<td>&gt;</td>
<td>S</td>
<td>W/</td>
</tr>
<tr>
<td>Mowry (1957)</td>
<td>18 harmonic</td>
<td>.72</td>
<td>SE</td>
<td>WC-S</td>
<td>--</td>
<td>S</td>
<td>W/</td>
</tr>
<tr>
<td>Argyle, Gardner,</td>
<td>87</td>
<td>.24</td>
<td>P</td>
<td>M</td>
<td>&lt;</td>
<td>S</td>
<td>W/O</td>
</tr>
</tbody>
</table>

---

a P = Productivity used as independent criterion  
SE = Supervisor effectiveness used as independent criterion
M = Manufacturing job site
BC-S = Blue collar - Service job site
WC-S = White collar - Service job site

<= Group size less than or equal to 10
> = Group size greater than 10
--- = Could not be calculated with provided information

S = Subjective measure used in obtaining independent variable
O = Objective measure used in obtaining independent variable

W/ = Study which had combined multiple effect sizes into one summary effect size
W/O = Study which had only one effect size
meta-analytic techniques (Hunter et al., 1982; Rosenthal, 1984; Glass et al., 1981). Therefore, the original statistical book, *Statistical Procedures and their Mathematical Bases*, (Peters and Van Voorhis, 1940) was obtained and the conversion of the epsilon statistic to an $F$ ratio was found. This $F$ ratio was then transformed into an $r$ value. This two step transformation is described in Appendix C, equations 1a and 1b.

When multiple effect sizes occurred in a study the conservative approach of combining the multiple effect sizes was taken. A method developed by Rosenthal and Rosen (1986), was used to calculate a combined effect size and protect the independence of the $rs$, as well as preventing overestimates of the significance of effects (Guzzo, Jette, & Katzell, 1985). The equations used and the values of the constants are described in Appendix C, equations 2a, 2b, and 2c.

**Cumulation of Effect Sizes**

The Hunter et al. (1982) method for cumulating effect sizes was used. This method allows for correcting of the error in sampling, the error of measurement, and restriction of range. Range restriction was not corrected for in this thesis project because the necessary information was absent. In addition, a ninety-five percent confidence interval was calculated for the mean effect
sizes of the total sample and all the subsamples. The equations used are presented in Appendix C, equations 3a through 3h. A chi-square test for heterogeneity of variance (Marascuilo, 1971) was conducted when unexplained variance remained after corrections. This is the same test used by other meta-analysis researchers (Fisher and Gitelson, 1983; Scott and Taylor, 1985). As a reviewer noted in the Fisher and Gitelson (1983, p.325) article, "this test yields a more accurate approximation to chi-square than the similar test given by Hunter, Schmidt, and Jackson (1982)." In fact, Hunter and his colleagues (1982) admit that their test is so powerful that it may identify even trivial amounts of unexplained variance as significant.

When the amount of unexplained variance was found to be significant, the studies were separated into subgroups based on potential moderator variables. A chi-square test for heterogeneity of variance was again conducted as well as a test of contrasts between the relevant subgroups. The equations for these tests are listed in Appendix C, equations 4a and 4b.

A BASIC program was written by the researcher which performs the majority of the calculations described in this section. Only the corrections for measurement error and the contrast test between relevant subgroups were not
included in the program. The program is listed in Appendix D.
RESULTS

The analysis of the data was originally conducted on the full set of 20 correlation coefficients. Table 2 lists the results of the initial analysis. Column 1 shows the particular variable of interest. The total sample size is given in column 2, and column 3 provides the number of effect sizes included in the analysis. The mean correlation weighted by sample size is provided in column 4, and column 5 shows the total observed variance in the sample correlations. The sampling error variance is contained in column 6, and the measurement variance is found in column 7. The value in column 8 is the unexplained variance, which is the difference of the sample variance, the sampling error variance, and the measurement variance. The final column is the results of the chi-square approximation test used to determine the significance of the unexplained error. The lower portion of Table 2 provides the corrected mean effect size across studies, the corrected variance across studies, the corrected standard deviation across studies, and the 95% confidence limits for the corrected mean effect size.

It can be seen from the overall analysis of the full
dataset that Hawthorne styles of supervision have a positive effect on productivity as well as perceived supervisor effectiveness. However, the 95% confidence limits range from -.06 to .62 which means that there is a possibility that the actual value could be near zero.

Table 2 shows that a large amount of the variance in the scores is still unexplained. This is supported by the significant chi-square value in column 9. These findings would seem to indicate that there are some variables which are moderating the relationship between Hawthorne styles of supervision and the independent criterion. Hunter et al. (1982) state that a moderator variable will show itself in two ways: (1) the average correlation will vary from subset to subset, and (2) the average unexplained variance will be lower in the subsets than for the data as a whole.

The first moderator variable examined was whether the independent criterion was productivity or supervisor effectiveness. The results of this meta-analysis are provided in Table 3. The two mean effect sizes are different from each other (.26 vs .41) and the amounts of unexplained variance in the subgroups are less than that of the data as a whole. However, supervisor effectiveness seems to be more of a moderator in that its mean effect size is quite different from the mean effect size observed in the complete data set and very little of the variance is
Table 2

Results of analysis of all the studies examining Hawthorne styles of supervision

<table>
<thead>
<tr>
<th>Total Sample Size</th>
<th>Number of Effect sizes</th>
<th>Mean Effect Size</th>
<th>Sampling Error Variance</th>
<th>Measurement Variance</th>
<th>Unexplained Variance</th>
<th>Chi-Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hawthorne styles of supervision</td>
<td>6813</td>
<td>20</td>
<td>.28</td>
<td>.02899</td>
<td>.00249</td>
<td>.00913</td>
</tr>
</tbody>
</table>

Note: Values in brackets are the percentages of the total variance which is being accounted for.

Corrected Mean Effect Size Across Studies
Corrected Variance Across Studies
Corrected Standard Deviation Across Studies
95% Confidence Limits

Hawthorne Styles of Supervision

\[
\text{Corrected Mean Effect Size Across Studies} = 0.37
\]

\[
\text{Corrected Variance Across Studies} = 0.02968
\]

\[
\text{Corrected Standard Deviation Across Studies} = 0.17228
\]

\[
95\% \text{ Confidence Limits} = -0.06 \leq \mu \leq 0.62
\]
Table 3

Results of analyses of studies using production as the independent criterion and those using supervisor effectiveness as the independent criterion

<table>
<thead>
<tr>
<th>Total Sample Size</th>
<th>Number of effect sizes</th>
<th>Mean Effect Size</th>
<th>Sample Variance</th>
<th>Sampling Error Variance</th>
<th>Measurement Variance</th>
<th>Unexplained Variance</th>
<th>Chi-Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Production as the criterion</td>
<td>5937</td>
<td>12</td>
<td>.26</td>
<td>.02664</td>
<td>.00175</td>
<td>.00771</td>
<td>.01718</td>
</tr>
<tr>
<td>Supervisor effectiveness as the independent criterion</td>
<td>876</td>
<td>8</td>
<td>.41</td>
<td>.02595</td>
<td>.00632</td>
<td>.01944</td>
<td>.00019</td>
</tr>
</tbody>
</table>

Note: Values in brackets are the percentages of the total variance which is being accounted for.

Corrected Mean Effect Size Across Studies
Production: .34
Supervisor Effectiveness: .54

Corrected Variance Across Studies
Production: .03236
Supervisor Effectiveness: .01111

Corrected Standard Deviation Across Studies
Production: .17988
Supervisor Effectiveness: .10543

95% Confidence Limits
Production: -.01 ≤ μ ≤ .69
Supervisor Effectiveness: .33 ≤ μ ≤ .75
left unexplained (0.7%).

The next moderator variable examined was the type of job site at which the study was conducted (manufacturing, white collar-service, and blue collar-service). These findings are contained in Table 4. The results show that all the mean effect sizes are different from each other but only studies conducted at a manufacturing site and those done at a blue collar-service site have unexplained variance less than that of the data as a whole. This suggests that the moderator white collar-service site has a submoderator variable(s) which is/are influencing its effect on the independent criterion.

Another moderator variable examined was those studies conducted during the 1950s as opposed to those conducted after the 1950s. Table 5 gives the findings to this meta-analysis, and the results show that this factor is a moderator as well. The mean effect sizes are different from one another and the amount of unexplained variance is less that that of the total sample. However, there still remains a significant portion of unexplained variance in both categories, which leads one to conclude that the two categories are not strong moderators.

Table 6 provides the results of the moderator variable of work group size. Only 11 studies provided sufficient information to code for this variable. The median work
Table 4

Results of analyses of studies from manufacturing sites, white collar service sites, and blue collar service sites

<table>
<thead>
<tr>
<th></th>
<th>Total Sample Size</th>
<th>Number of effect sizes</th>
<th>Mean Effect Size</th>
<th>Sampling Variance</th>
<th>Measurement Variance</th>
<th>Unexplained Variance</th>
<th>Chi-Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manufacturing site</td>
<td>1527</td>
<td>7</td>
<td>.30</td>
<td>.00241</td>
<td>.00377</td>
<td>.01014</td>
<td>.00</td>
</tr>
<tr>
<td>Blue collar service site</td>
<td>3876</td>
<td>4</td>
<td>.19</td>
<td>.00765</td>
<td>.00096</td>
<td>.00417</td>
<td>.00252</td>
</tr>
<tr>
<td>White collar service site</td>
<td>1410</td>
<td>9</td>
<td>.49</td>
<td>.05007</td>
<td>.00365</td>
<td>.02731</td>
<td>.01911</td>
</tr>
</tbody>
</table>

Note: Values in brackets are the percentages of the total variance which is being accounted for.

Corrected Mean Effect Size Across Studies:
- Manufacturing: .39
- Blue Collar Service: .25
- White Collar Service: .64

Corrected Variance Across Studies:
- Manufacturing: -.01965
- Blue Collar Service: .00431
- White Collar Service: .03266

Corrected Standard Deviation Across Studies:
- Manufacturing: 0.0
- Blue Collar Service: .06562
- White Collar Service: .18073

95% Confidence Limits:
- Manufacturing: \( \mu = 0.39 \)
- Blue Collar Service: \( 0.12 \leq \mu \leq 0.38 \)
- White Collar Service: \( 0.29 \leq \mu \leq 0.99 \)
Table 5

Results of analyses of studies conducted in the 1950s and those conducted after the 1950s

<table>
<thead>
<tr>
<th></th>
<th>Total Sample Size</th>
<th>Number of Effect sizes</th>
<th>Mean Effect Size</th>
<th>Sampling Error Variance</th>
<th>Measurement Variance</th>
<th>Unexplained Variance</th>
<th>Chi-Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>During the 1950s</td>
<td>2051</td>
<td>12</td>
<td>.34</td>
<td>.00460</td>
<td>.01951</td>
<td>.01291</td>
<td>164.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(12.4%)</td>
<td>(52.7%)</td>
<td>(34.9%)</td>
</tr>
<tr>
<td>After the 1950s</td>
<td>4762</td>
<td>8</td>
<td>.26</td>
<td>.00145</td>
<td>.00771</td>
<td>.01446</td>
<td>146.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(6.1%)</td>
<td>(32.6%)</td>
<td>(61.2%)</td>
</tr>
</tbody>
</table>

Note: Values in brackets are the percentages of the total variance which is being accounted for.

Corrected Mean Effect Size Across Studies
- During the 1950s: .44
- After the 1950s: .34

Corrected Variance Across Studies
- During the 1950s: .03334
- After the 1950s: .02468

Corrected Standard Deviation Across Studies
- During the 1950s: .18260
- After the 1950s: .15711

95% Confidence Limits
- During: .08 < μ < .80
- After: .03 < μ < .65
Table 6

Results of analyses of studies with an average work group size less than or equal to 10 and those with an average work group size greater than 10

<table>
<thead>
<tr>
<th></th>
<th>Total Sample Size</th>
<th>Number of effect sizes</th>
<th>Mean Effect Size</th>
<th>Sampling Variance</th>
<th>Measurement Variance</th>
<th>Unexplained Variance</th>
<th>Chi-Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group size less than or equal to 10</td>
<td>577</td>
<td>5</td>
<td>.21</td>
<td>.01824</td>
<td>.00791</td>
<td>.00739</td>
<td>12.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(.43.4%)</td>
<td>(16.1%)</td>
<td>(40.5%)</td>
<td>p &lt; .05</td>
</tr>
<tr>
<td>Group size greater than 10</td>
<td>5107</td>
<td>6</td>
<td>.25</td>
<td>.02260</td>
<td>.00104</td>
<td>.01430</td>
<td>219.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(.4.6%)</td>
<td>(32.1%)</td>
<td>(63.3%)</td>
<td>p &lt; .01</td>
</tr>
</tbody>
</table>

Note: Values in brackets are the percentages of the total variance which is being accounted for.

<table>
<thead>
<tr>
<th></th>
<th>Group size ≤ 10</th>
<th>Group size &gt; 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Mean Effect Size Across Studies</td>
<td>.27</td>
<td>.33</td>
</tr>
<tr>
<td>Corrected Variance Across Studies</td>
<td>.01263</td>
<td>.02444</td>
</tr>
<tr>
<td>Corrected Standard Deviation Across Studies</td>
<td>.11240</td>
<td>.15633</td>
</tr>
<tr>
<td>95% Confidence Limits</td>
<td>.05 ≤ μ ≤ .49</td>
<td>.02 ≤ μ ≤ .64</td>
</tr>
</tbody>
</table>
group size was 10, which is why that particular cutoff point was used. Both groups met the criterion of a moderator variable with work groups of less than 10 having more of a moderating effect on the independent criterion because of its greater accounting of variance.

Those moderator variables which are based on the methodological characteristics of the studies are examined next. The effect of the sample size (<100 vs >100) on the independent criterion seems to be minimal. The correlations of the two subgroups are almost equal and one of the subgroups has a larger amount of unexplained variance that the population as a whole. An interesting side note is the size of the sampling error variance for the two groups. Those studies with less than 100 subjects had about 50% of their variance being accounted for by sampling error, while studies with over 100 subjects have only a minimal amount of sampling error variance. Although this is the expected finding, it helps illustrate the need for adequate sampling when conducting experiments.

Another methodological moderator examined was whether an objective or a subjective independent criterion measure was used. Table 8 shows the findings from this meta-analysis. The results support this variable as a moderator of Hawthorne styles of supervision. The two mean correlations are widely separated (.18 vs .38) and the
Table 7
Results of analyses of studies with sample sizes less than 100 and those with sample sizes greater than 100

<table>
<thead>
<tr>
<th>Sample size less than 100</th>
<th>Total Sample Size</th>
<th>Number of effect sizes</th>
<th>Mean Effect Size</th>
<th>Sampling Error Variance</th>
<th>Measurement Variance</th>
<th>Unexplained Variance</th>
<th>Chi-Square</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>526</td>
<td>10</td>
<td>.30</td>
<td>.03121</td>
<td>.01571</td>
<td>.01014</td>
<td>.00536</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(50.3%)</td>
<td>(32.5%)</td>
<td>(17.2%)</td>
<td>p &lt; .01</td>
</tr>
<tr>
<td>Sample size greater than 100</td>
<td>6287</td>
<td>10</td>
<td>.28</td>
<td>.02876</td>
<td>.00135</td>
<td>.00913</td>
<td>.01828</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(4.7%)</td>
<td>(31.7%)</td>
<td>(63.6%)</td>
<td>p &lt; .01</td>
</tr>
</tbody>
</table>

Note: Values in brackets are the percentages of the total variance which is being accounted for.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 100</td>
<td>&gt; 100</td>
</tr>
<tr>
<td>Corrected Mean Effect Size Across Studies</td>
<td>.39</td>
</tr>
<tr>
<td>Corrected Variance Across Studies</td>
<td>.00916</td>
</tr>
<tr>
<td>Corrected Standard Deviation Across Studies</td>
<td>.09573</td>
</tr>
<tr>
<td>95% Confidence Limits</td>
<td>.20 ≤ μ ≤ .58</td>
</tr>
</tbody>
</table>
Table 8

Results of analyses of studies with objective independent criteria and those with subjective independent criteria

<table>
<thead>
<tr>
<th></th>
<th>Total Sample Size</th>
<th>Number of effect sizes</th>
<th>Mean Effect Size</th>
<th>Sample Variance</th>
<th>Sampling Error Variance</th>
<th>Measurement Variance</th>
<th>Unexplained Variance</th>
<th>Chi-Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective independent criterion</td>
<td>3410</td>
<td>6</td>
<td>.18</td>
<td>.00796</td>
<td>.00165</td>
<td>.00384</td>
<td>.00247</td>
<td>33.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p &lt; .01</td>
</tr>
<tr>
<td>Subjective independent criterion</td>
<td>3403</td>
<td>14</td>
<td>.38</td>
<td>.02909</td>
<td>.00299</td>
<td>.01667</td>
<td>.00943</td>
<td>193.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p &lt; .01</td>
</tr>
</tbody>
</table>

Note: Values in brackets are the percentages of the total variance which is being accounted for.

Corrected Mean Effect Size Across Studies

<table>
<thead>
<tr>
<th></th>
<th>Objective</th>
<th>Subjective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Mean Effect Size Across Studies</td>
<td>.24</td>
<td>.50</td>
</tr>
<tr>
<td>Corrected Variance Across Studies</td>
<td>.00422</td>
<td>.01612</td>
</tr>
<tr>
<td>Corrected Standard Deviation Across Studies</td>
<td>.06498</td>
<td>.12696</td>
</tr>
<tr>
<td>95% Confidence Limits</td>
<td>.11 ≤ μ ≤ .37</td>
<td>.25 ≤ μ ≤ .75</td>
</tr>
</tbody>
</table>
unexplained variance is less than that of the overall population. However, there still remains a significant amount of unexplained variance which leads one to believe that this variable is not a moderator by itself.

The last moderator examined is contained in Table 9. This analysis was to check to see if there was a significant difference in studies with a combined effect size as opposed to studies without a combined effect size. The findings do not support this variable as a moderator. The mean correlations, although different, are only slightly so (.27 vs .30). In addition, one of the groups had more unexplained variance than the data as a whole. This leads one to reject combined versus uncombined effect sizes as an influence on the conducted meta-analysis.

Contrast tests were conducted to examine any significant differences between the effect sizes of the moderating subgroups. Three comparisons resulted in significant differences. The moderating groups of manufacturing site and white collar-service ($z = 4.61, p < .0000, 2$-tailed), blue collar-service and white collar-service ($z = 6.63, p < .0000$), and work groups of 10 or less and work groups greater than 10 ($z = 1.98, p = .0478, 2$-tailed) were found to differ significantly.

The findings of this study are based on the assumption that the literature review has found all relevant studies.
Table 9
Results of analyses of studies without a combined effect size and those with a combined effect size

<table>
<thead>
<tr>
<th>Total Sample Size</th>
<th>Number of Effect Sizes</th>
<th>Mean Effect Size</th>
<th>Sample Error Variance</th>
<th>Measurement Variance</th>
<th>Unexplained Variance</th>
<th>Chi-Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Studies without a combined effect size</td>
<td>3882</td>
<td>11</td>
<td>.27</td>
<td>.03667</td>
<td>.00244</td>
<td>.00817</td>
</tr>
<tr>
<td>(6.7%)</td>
<td>(22.3%)</td>
<td>(71.1%)</td>
<td>p &lt; .01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Studies with a combined effect size</td>
<td>2931</td>
<td>9</td>
<td>.30</td>
<td>.01842</td>
<td>.00255</td>
<td>.01014</td>
</tr>
<tr>
<td>(13.8%)</td>
<td>(55.0%)</td>
<td>(31.1%)</td>
<td>p &lt; .01</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Values in brackets are the percentages of the total variance which is being accounted for.

Corrected Mean Effect Size Across Studies
Without Combined .35
Corrected Variance Across Studies
Without Combined .04454 .00978
Corrected Standard Deviation Across Studies
Without Combined .21104 .09887
95% Confidence Limits
Without Combined -.06 < μ ≤ .76 .20 < μ ≤ .58
This does not take into account those studies which could have had null results and were not published in a journal. To examine the potential for this problem, Rosenthal's "file drawer" equation was used (Rosenthal, 1984) to calculate the number of studies averaging null results which would be required to reduce the findings of this study to nonsignificance. The results showed that 3544 studies with null results would be needed to reverse the findings of the overall meta-analysis.
DISCUSSION

The Hawthorne Findings

The conclusions reached by the Hawthorne experimenters were revolutionary when they were first published. The Hawthorne studies illustrated how workers were not automatons who gave constant levels of output but who were effected by human emotions and motivational factors. These findings were the origin of the 'human relations' influence on U.S. management styles. The conclusions of the researchers at Hawthorne were not without their critics. These criticisms ranged from critiques of the methodology employed to replacement of the original conclusions with new ones based on operant conditioning, learning models, etc. The purpose of this study was to examine those experiments subsequent to the publication of the Hawthorne findings to clarify the effects of friendly supervision on worker production or supervisor effectiveness as perceived by management.

The initial analysis on the 20 studies seems to have confirmed the impact of Hawthorne styles of supervision on productivity/supervisor effectiveness. However, given the 95% confidence interval listed in Table 2, the possibility exists that friendly supervision may have no effect on
production/supervisor effectiveness. In addition, when only those studies using productivity as the independent criterion were examined (these studies most closely resembling the conditions of the Hawthorne studies) it was found that the corrected mean effect size was only slightly smaller than that of the complete set of studies (all 20 studies $\bar{r} = .37$; studies which used productivity $\bar{r} = .34$). Again, this seems to imply that Hawthorne styles of supervision do positively influence production, which is the "bottom line" for many organizations. However, the 95% confidence interval (shown in Table 2) does suggest that friendly supervision has the possibility of not effecting production.

The findings described above can be viewed as a confirmation for both the original Hawthorne studies and the criticisms of the Hawthorne studies. This is because friendly supervision does have a moderately positive effect on production, but there remains the possibility that friendly supervision has no relationship with production. The findings of this thesis project would then seem to leave the conclusions of the Hawthorne researchers and their critics as muddled as ever. However, this is not the case.

This thesis project has clarified many aspects of the relationship between friendly supervision and
productivity/supervisor effectiveness. One of the first areas of interests is the type of independent criterion used—productivity as opposed to supervisor effectiveness. It was found that supervisor effectiveness had a larger corrected mean effect size when compared to that of productivity (supervisor effectiveness $\bar{F} = .54$; productivity $\bar{F} = .34$). If one assumes that those studies which used supervisor effectiveness as the independent criterion had the same productivity levels as those studies using productivity as the independent criterion, then this suggests that supervisors who use a Hawthorne style of supervision are rated as being more effective by management even though the productivity levels may have had little or no relationship with the supervisors' styles. An implication of this finding is that supervisors are evaluated on factors which are not entirely dependent on production rates. The process model of performance rating developed by Landy and Farr (1980) supports this finding.

Another aspect of this study was the finding of how the site of the study (manufacturing, blue collar-service, and white collar-service) demonstrated how the effectiveness of a style of supervision was dependent on the type of organization. White collar-service organizations were highly effected by Hawthorne styles of supervision, while blue collar-service and manufacturing were influenced to a
lesser degree. The difference between white collar-service and the other types of organizations were statistically significant as well.

This finding has an implication for future research. Specifically, one could examine how different styles of supervision, covering the spectrum of supervisory styles (i.e., autocratic to laissez faire), effect different job types. It could be possible that the optimum style of supervision is not the same for different job types. In addition, the different supervisory styles and job types could be examined while taking additional factors such as work group size, number of supervisors in the organizational hierarchy, or the level of unionization, into account.

The finding that work groups of 10 or less members had smaller corrected mean effect sizes than work groups of more than 10 workers ($< 10 \overline{r} = .27; > 10 \overline{r} = .33$) goes against what was expected. Porter and Lawler's (1965) study examining organizational subunit size and job behaviors suggests that the relationship is negative or curvilinear. However, the findings of this thesis project were based on only 11 studies with a median work group size of 10. This is a rather limited sample given the range of the sizes of work groups in organizations, and therefore no definite conclusions concerning work group size can be
drawn from the current study.

When the independent criterion was objective the corrected mean effect size was smaller ($\bar{F} = .24$) than when compared to a subjective independent criterion ($\bar{F} = .50$). Although these two effect sizes were not significantly different, the effect sizes suggest that objective and subjective measures may not be measuring the same constructs, or if they are, they are doing so differently. This implies that rating bias is involved in the subjective measures of production and supervisor effectiveness.

The Meta-Analytic Method

Meta-analysis is becoming an established method of reviewing research in the social sciences. There are problems with the technique, but overall, meta-analysis is the most objective and thorough literature review method available. This study used the meta-analytic technique developed by Hunter and his colleagues. This method is known for its corrections of the following possible errors: (1) sampling error, (2) measurement error, and (3) range restriction. Other meta-analytic techniques exist (cf. Glass et al., 1981; Hedges and Olkin, 1982; Rosenthal, 1984) but at the time of this study there was no clear consensus for a preferred method. This lack of consensus should not be viewed as evidence of an inherent weakness of meta-analysis but rather as the refinement of a new
scientific instrument.

However, for the researcher attempting to make sense of an area of interest by using meta-analysis, these differences can be a cause of problems and concerns about the validity and future acceptance of the results of a meta-analysis in progress. Studies using Monte Carlo methods are currently being done in an attempt to find the strengths and weaknesses of the meta-analytic techniques. Hopefully, there will soon be an accepted and proven method with which to conduct a quantitative literature review.

Those who criticize meta-analysis on grounds of sample bias, improper coding of data, and biased interpretation do not appreciate some of the finer qualities of the meta-analytic method. Not only does a meta-analysis review the literature with greater precision than is normally accomplished, but it allows scrutiny and criticism of the methods used by the person conducting the meta-analysis. If something is found to be less-than-perfect about a meta-analysis, it is most likely due to the user of the technique and not the technique itself.

Bangert-Drowns (1986) summarizes the meta-analytic method this way:

Meta-analysis is not a fad. It is rooted in the fundamental values of the scientific enterprise: replicability, quantification, causal and correlational analysis. Valuable information is needlessly scattered in individual studies. The ability of social scientists to
deliver generalizable answers to basic questions of policy is too serious a concern to allow us to treat research integration lightly. The potential benefits of meta-analysis method seem enormous. (p. 276)

Summary

This thesis project was an attempt to confirm or refute the conclusions of the Hawthorne researchers. It involved the use of a quantitative literature review method known as meta-analysis, and was conducted with 20 studies which examined aspects of styles of supervision at an actual job site. The findings indicate that there is a positive relation between friendly styles of supervision and productivity levels or rated supervisor effectiveness. However, the lower limit of the 95% confidence interval was found to be slightly negative so the possibility exists that friendly styles of supervision have little or no effect on the independent criterion.
APPENDIX A

Earlier Criticisms of Hawthorne Research


APPENDIX B

Reference List of Examined Articles


Gutkin, T. B. (1978). Modification of elementary students'


Sorensen, L. R. (1967). A study of supervisory leadership behavior, technical competence, and group
effectiveness. Dissertation Abstracts, 28(3-B), 1252.


APPENDIX C

Transformation of epsilon to an effect size

Peters and Van Voorhis (1940, p. 353) list the equation of transforming epsilon to an F-ratio as:

\[ F = \frac{(N-k)e^2 + (k-1)}{(k-1)(1-e^2)} \]

where \( N \) is the size of the sample, and \( k \) is the number of groups.

The resulting \( F \) was then transformed to an \( r \) value using equation 2.17 in Rosenthal (1984):

\[ r = \sqrt{\frac{F(1,-)}{F(1,-) + df\ error}} \]

where \( F(1,-) \) indicates any \( F \) with \( df = 1 \) in the numerator, and the dash represents \( df \) in the denominator.

Combining multiple effect sizes in a study

The equation for a combined or composite effect size, \( e_c \) is found in Rosenthal and Rubin (1986, p. 402 as:

\[ e_c = \frac{\Sigma \lambda_i t_i / I}{\{\rho(\Sigma \lambda_i)^2 + (1-\rho)\Sigma \lambda_i^2\}} \]

where \( I \) is the index of the size of the study, in this

92
study I was defined as \(((n-1)/2)^{1/2}\); \(\lambda_i\) is the weight assigned to the ith dependent variable. Equal weights of 1 were used for this study. \(\rho\) is the typical intercorrelation among the dependent variables. The mean of those studies listing intercorrelations was .45 in this study. \(t_i\) is defined as:

\[ t_i = \frac{r_i (df)^{1/2}}{(1-r_i^2)^{1/2}} \]

The composite effect size was then expressed as a correlation by the equation:

\[ r_c = \frac{e_c}{\{e_c^2 + (df)/I^2\}^{1/2}} \]

where the df was N-2. When unequal sample sizes were encountered the harmonic mean was used in the calculation of both the I and df values.

**Cumulation of effect sizes**

The equation for calculating the mean population correlation, is from Hunter et al. (1982):

\[ \bar{r} = \frac{\sum(N_i r_i)}{\sum N_i} \]

where \(N_i\) and \(r_i\) are the sample size and effect size for the ith study. This value is then used to calculate the variance across studies (Hunter et al., p. 41) which
is the frequency weighted average squared error:

(3b) \[ \sigma^2_r = \frac{\sum N_i (r_i - \bar{r})^2}{\sum N_i} \]

The error variance across studies is the average within study variance given by the "almost perfect approximation" (Hunter et al., 1982, p. 44):

(3c) \[ \sigma^2_e = \frac{(1-\bar{r}^2)^2 K}{N} \]

where \( K \) is the number of studies and \( N = \sum N_i \) or the total sample size.

The variance due to measurement error is calculated by the following equation (Hunter et al., 1982, p. 83):

(3d) \[ \sigma^2_{ME} = \bar{\rho}^2_{TU} (\bar{a}^2 \sigma^2_a + \bar{b}^2 \sigma^2_b) \]

where \( \bar{\rho}_{TU} = \frac{\bar{\rho}_{XY}}{\bar{a} \bar{b}} \); and \( \sigma^2_b = \frac{\Sigma (r_{YY_i} - \bar{b})^2}{\sum r_{YY_i}} \)

where \( r_{xx} \) is the reliability of the independent variables and \( r_{yy} \) is the reliability of the dependent variables. \( a \) is the square root of \( r_{xx} \) and \( b \) is the square root of \( r_{yy} \).

The true score mean correlation is found by calculating (Hunter et al., 1982, p. 80):
3e) \[
\overline{\rho}_{TU} = \frac{\rho_{xy}}{ab} = \frac{r}{ab}
\]

The true score variance is found from (Hunter et al., 1982, p. 80):

3f) \[
\sigma^{2}_{\rho_{xy}} = \overline{\rho}_{TU}^2 \left( \overline{a^2\sigma^2} + \overline{B^2\sigma^2} \right)
\]

The standard deviation of the true scores is:

3g) \[
\sigma_{\rho_{TU}} = \left( \sigma^{2}_{\rho_{TU}} \right)^{1/2}
\]

The standard equation for calculating confidence intervals was used:

3h) \[
\overline{\rho}_{TU} - 1.96\sigma_{\rho_{TU}} \leq \mu \leq \overline{\rho}_{TU} + 1.96\sigma_{\rho_{TU}}
\]

**Test for heterogeneity of variance**

The chi-square test to test for heterogeneity of variance was calculated by the following equation (Marascuilo, 1971, p. 452):

4a) \[
U_0 = \sum_{k=1}^{k} \left( N_k - 3 \right)(z_k - z_0)^2
\]

approximates \( \chi^2 \) with \( k-1 \) degrees of freedom.

where \( k \) is the number of correlations (transformed
into Fisher's z-scores); \( N_k \) is the number of subjects in the kth study; \( z_k \) is the z value for the kth study and \( z_0 \) is the mean z score given by:

\[
z_0 = \frac{\sum_{k=1}^{k} (N_k - 3)z_k}{\sum_{k=1}^{k} (N_k - 3)}
\]

Post hoc contrast test

The statistical significance of the contrast, testing any specific hypothesis about the set of effect sizes, can be obtained from a \( Z \) computed as follows (Rosenthal, 1984, p. 84):

\[
4b) \quad \frac{\Sigma \lambda_j z_{r_j}}{\Sigma \lambda_j^2 w_j}
\]

where \( \lambda_j \) is the contrast weight determined from some theory for any one study, such that the sum of \( \lambda_j = 0 \); \( z_{r_j} \) is the Fisher's \( z_r \) for any one study; \( w_j \) is the inverse of the variance of the effect size for each study. For Fisher's \( z_r \) transformations of the effect size \( r \), the variance is \( 1/(N_j - 3) \) so \( w_j \) is \( N_j - 3 \).
APPENDIX D

BASIC Program

The format of the data file is one value per line. The first two lines are: #1 Number of effect sizes in the data file; #2 Descriptive name of the data set. The values to be analyzed then follow, with one data value per line. The program was written for a Radio Shack TRS-80 64K Color Computer 2.

CLEAR 1500
D=0:E=0:SM=0:SN=0:SV=0:XR=0:SNZ=0:SZ=0
EV=0:V=0:STD=0:XSZ=0:ZV=0:SMM=0
CLS
INPUT"NAME OF DATA FILE";G$
INPUT"NAME OF OUTPUT FILE";H$
CLS
PRINT"PUT IN DATA CASSETTE AND"
PRINT"POSITION DATA FILE"
PRINT"PRESS <PLAY>"
INPUT"PRESS <ENTER> TO CONTINUE";T$
OPEN "I",#@-1, G$
IF EOF (-1) THEN END
INPUT #@-1, A, B$
DIM I(A);DIM R(A);DIM N(A);DIM Z(A);DIM NZ(A)
DIM IS(A);DIM R$(A);DIM N$(A);DIM Z$(A);DIM NZ$(A)
DIM SSM(A)
FOR X=1 TO A
IF EOF THEN END
INPUT #@-1, I(X), R(X), N(X)
I$(X)=STR$(I(X));R$(X)=STR$(R(X)); N$(X)=STR$(N(X))
NEXT X
CLOSE #@-1
PRINT"PUT IN OUTPUT CASSETTE AND"
PRINT"POSITION TAPE"
PRINT"PRESS <PLAY-RECORD>"
INPUT"PRESS <ENTER> TO CONTINUE";S$
OPEN "O", #@-1, H$
PRINT #@-1;PRINT #@-1;PRINT #@-1
PRINT -1, A$ + B$
PRINT -1: PRINT -1: PRINT -1
PRINT -1, "INPUTTED DATA": PRINT -1
FOR X = 1 TO A
E$ = A$ + "ID NUMBER" + I$(X) + A$ + "EFFECT SIZE" + R$(X) + A$ + "SAMPLE SIZE" + N$(X)
PRINT -1, E$: PRINT -1
NEXT X
PRINT -1: PRINT -1
PRINT -1, "PRODUCTS OF EFFECT SIZE* N"
FOR X = 1 TO A
SMM(X) = N(X) * R(X)
SM = SMM(X) + SM
SN = N(X) + SN
SM$ = STR$(SMM(X))
PRINT -1, A$ + I$(X) + A$ + SM$
NEXT X
SM$ = STR$(SM)
SN$ = STR$(SN)
XR = SM / SN
XR$ = STR$(XR)
PRINT -1, A$ + SM$ + "/" + SN$
FOR X = 1 TO A
D = R(X) - XR
E = D**2
SV = (N(X) * E) + SV
NEXT X
RV = SV / SN
RV$ = STR$(RV)
EV = (A*((1-XR**2)**2))/SN
EV$ = STR$(EV)
V = RV - EV
V$ = STR$(V)
STD = SQR(V)
STD$ = STR$(STD)
PRINT -1
E$ = A$ + "MEAN EFFECT SIZE" + XR$
PRINT -1, E$
PRINT -1
E$ = A$ + "VARIANCE IN EFFECT SIZE" + RV$
PRINT -1, E$
PRINT -1
E$ = A$ + "VARIANCE DUE TO SAMPLING ERROR" + EV$
PRINT -1, E$
PRINT -1
E$ = A$ + "CORRECTED VARIANCE IN EFFECT SIZE" + V$
PRINT -1, E$
PRINT -1
E$ = A$ + "STANDARD DEVIATION IN EFFECT SIZE" + STD$
PRINT #-1, E$
SNZ=0
FOR X=1 TO A
  Z(X) = .5*LOG((1+R(X))/(1-R(X)))
  Z$(X)=STR$(Z(X))
  NZ(X)=N(X)-3
  NZ$(X)=STR$(NZ(X))
  SNZ=NZ(X)+SNZ
  SNZ$=STR$(SNZ)
  SZ=(Z(X)*NZ(X))+SZ
  SZ$=STR$(SZ)
PRINT #-1
  E$= A$+"ID NUMBER" + I$(X) + A$ + "FISHER Z VALUE" + Z$(X)
PRINT #-1, E$ + A$ + "VAR 1/(N-3) " + NZ$(X)
NEXT X
XSZ=SZ/SNZ
XSZ$=STR$(XSZ)
PRINT #-1:PRINT #-1
PRINT #-1, A$+"SQUARED DIFFERENCES OF Z(X) AND MEAN Z"
FOR X=1 TO A
  PZX=(Z(X)-XSZ)**2
  E$=STR$(PZX)
PRINT #-1
  E$= A$+I$(X) + NZ$(X) + E$
  ZV=NZ(X)*((Z(X)-XSZ)**2)+ZV
NEXT X
  ZV$=STR$(ZV)
  SZ$=STR$(SZ)
  SNZ$=STR$(SNZ)
PRINT #-1
PRINT #-1, A$ + "SUM OF DIFFERENCES " + SZ$ + A$ + "SUM OF N " +SNZ$
PRINT #-1
  E$=A$+"MEAN FISHER Z" + XSZ$
PRINT #-1, E$
PRINT #-1
  E$=A$+"CHI-SQUARE SCORE" + ZV$
PRINT #-1, E$
  DF$=STR$(A-1)
  Y$=STRING$(3,95)
PRINT #-1
  E$= A$+ "FOR ALPHA = .05, CHI-SQUARE("+DF$+") =" 
PRINT #-1, E$+Y$
CLOSE #-1
PRINT"DO YOU WISH TO RUN ANOTHER DATA"
PRINT"FILE?  Y/N"
INPUT C$
IF C$="Y" THEN 100 ELSE END
END
REFERENCES


American Psychologist, 37, 856.


