

Chapter 3

Methodology

This chapter presents the methodology used in the research synthesis of quantitative studies that address the influence of school leadership on the performance level of the school. The steps adopted for the conduct of the study included: locating, selecting, and evaluating the studies along predetermined criteria; identifying and coding variables; classifying the studies within a specific framework; calculating a common metric; aggregating the data; performing the meta-analysis; and analyzing the data. Each step is discussed in detail in this chapter.

Meta-analysis

Meta-analysis is a statistical analytical approach utilized for the purpose of synthesizing quantitative data to determine the degree of effect of the independent variable on the dependent variable. It uses statistical results from individual quantitative studies as the units of measurement in an overall study. These results are summarized by indices of effect size or standardized mean difference that “may then be averaged to obtain an overall estimate of effect magnitude” (Hedges & Becker, 1986, p. 15). In this study, the effects of principal leadership on school effectiveness, as indicated by the results of individual studies, were examined to determine the size of the overall effect.

The strength of the meta-analysis derives from the fact that the indices of effect magnitude are scale-free. Particularly for the social sciences where constructs are measured on a number of different instruments (see Appendix A) that yield numbers on a completely different scale (Hedges & Olkin, 1985), the use of a scale-free index of effect magnitude makes it possible to combine the results of studies that measure the same construct.

Locating the Studies

A comprehensive attempt was made to identify relevant studies on principal leadership and school effectiveness that fell within the specific time frame of 1985 to 2000. Initial searches used the Educational Resources Information Center (ERIC) First Search on-line retrieval system to search the ERIC and Current Journals in Education (CJIE) databases. This was followed by a search of on-line journals accessed through Info-Trac, and abstracts of theses and dissertations accessed through the Dissertation Abstracts database.

The key descriptors used included: Leadership and school effectiveness; school leadership and student outcomes; effective school administration, and student achievement. The combined searches yielded over 600 citations for possible consideration. A careful on-line review of the abstracts reduced the number to less than 130 as most of the studies were either anecdotal in nature, did not address a relationship between the independent and dependent variables, or did not include sufficient statistical data to compute an effect size.

With retrieval of articles and dissertations, a branching search of the references was undertaken. This cross-referencing yielded an additional eight studies to be considered for possible inclusion in the research. A further review of the related literature identified no new studies that could be included in the quantifiable database of a meta-analysis.

Measuring Leadership and Effectiveness

As discussed in Chapter 2, there are many approaches to leadership and these are reflected in the various instruments used in measuring the construct. As can be seen in Appendix A, approximately 20 different instruments were used in the various studies either singly, or in combination, as a measure of leadership effectiveness, or student outcomes. The instruments used included the short form of the Leadership Behavior Descriptive Questionnaire

(LBDQ-X11) developed by Stogdill (1963) out of the Ohio State University Leadership Studies (Bass & Stogdill, 1990). This instrument has been used extensively in various settings and has reliabilities of .90 for the consideration dimension and .78 for initiating structure.

Another instrument used in measuring leadership was the Leadership Practices Inventory-Self (LPI-S) and Leadership Practices Inventory-Other (LPI-O). This instrument was developed by Kouzes and Posner and consists each of 30 behaviorally based statements. The overall reliability for the instrument is very high ($\alpha = .97$).

Other instruments of established reliability and validity used were Likert's (1972) Profile of a School Survey Staff Questionnaire (POS); reliability .70 -.90; the Principal Instructional Management Rating Scale (PIMRS) developed by Hallinger and Murphy (1985) and used as a measure of the instructional leadership of the principal; the Profile for Assessment of Leadership (PAL) developed by the Leadership Assessment Development Committee (1984); and the short form of the Multifactor Leadership Questionnaire (MLQ-5X) (Bass, 1985) used as a measure of transformational leadership. A full list of the instruments used can be found in Appendix A.

While the authors of most of the journal articles developed their own instruments, and in the case of the Australian studies, used a measure of transformational leadership adapted to that context, almost all the dissertations used an established instrument of measurement.

For the most part, school effectiveness was measured on student performance on a standardized test: for example, the Iowa Test of Basic Skills (ITBS), the California Assessment Test (CAT). However, as argued by Cheng (1994), Leithwood and Jantzi (1999), Silins, Mulford and Zarins (1999), the effectiveness of a school is reflected by much more than just academic performance. Therefore studies that measured effectiveness on variables that looked at other student outcomes were included. An example is the Leithwood and Jantzi (1999) study that

looked at student engagement with school. They used as a measure the Student Engagement and Family Culture Survey, reported reliability coefficients, $r = .74$ for the participation sub-scale, and $r = .90$ for identification with the school. As they assert, there is a modest amount of evidence that links student engagement with school to academic performance in social studies, math, and language achievement.

Criteria for Selection

The following criteria guided the selection of studies for review. According to Hedges and Becker (1986), to be able to infer from the collection of effect sizes, three conditions must be satisfied. These are: (1) The estimates must represent the same construct; (2) they must be independent, and (3) they must estimate the same statistical parameters. In ensuring that this basic criterion was satisfied, first, all studies had to have been experimental, quasi-experimental or correlational. In this meta-analysis, there were no experimental studies. The nature of leadership in the educational context makes it virtually impossible to have a random assignment to any specific model. Consequently, the located studies were either quasi-experimental, or correlational.

Second, the studies must have been designed explicitly to examine leader behavior, or leadership style at the school-site level. Thus, excluded were all studies from the organizational management literature that were not based on school personnel. Also, principal leadership had to have been clearly conceptualized and measured by a standardized instrument as the independent variable.

Third, all studies must have included as the dependent variable, an explicit measure by which school effectiveness was determined. Most of the studies used student academic achievement on some standardized tests as the measure of school performance. However, some

studies looked at alternative outcome variables as a measure of effectiveness. Therefore, considered for inclusion in the meta-analysis were studies that examined the principal's impact on other student level outcome variables such as attendance (Blank, 1987) and student engagement in school, (Leithwood, & Jantzi, 1999), and studies that reported on the overall effectiveness of the school, for example, Cheng (1991).

The importance of the mediating teacher and school level variables has been detailed in the effective schools literature (Chubb, 1987; Rutter, 1979). Therefore, in acknowledging the "importance of classroom and school-level variables in contributing to school effectiveness" (Hallinger & Heck, 1998, p. 161), included in this quantitative review were studies that examined the indirect effect of the principal on school effectiveness (e.g. Cheng, 1991; Ogawa & Hart, 1985).

Fourth, all studies had to have adequate statistical data that would allow for the calculation of the effect size. For those studies that conducted descriptive analyses, necessary statistics for the transformations into effect sizes would include means and standard deviations. For studies that used inferential statistics, the appropriate transformation equations were applied. All studies that used multiple regression, multivariate analyses, including canonical correlations as the methods of analyses, had to have reported the zero-order correlations to be included in the study (Hedges & Olkin, 1985; Hunter & Schmidt, 1990). Based on the above criteria, 147 studies were excluded, leaving 47 studies for the analysis.

Further analysis during the encoding process resulted in seven studies being excluded from the final calculation. All seven studies did not allow for the calculation of a common metric. One study was qualitative but the style of presentation did not allow ready discernment of this fact, another focused on the nature of the principal's job, yet another examined only the

principal's impact on the organizational climate and teacher effectiveness, and the other four did not include sufficient statistical information for calculating the effect size.

Of the four that lacked statistical data, two reported inferential statistics and two were correlational. One study reported the F values, the degrees of freedom and the N . However, while this information may be sufficient for an understanding of the findings, and may have satisfied the requirements of the journal, they did not allow for the calculation of the d . For an ANOVA with $K > 2$ where K is the number of groups, it is necessary to report the means with the F - values. This allows for the calculation of the MS_w and the application of the formula

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{MS_w}}$$

(Hattie, 1998).

In the absence of reporting the means, at a minimum, the sums of squares (SS) and the degrees of freedom must be reported so that the one-way design could be approximated by adding into the error term the between-group sums of squares, except the SS of the variable of interest (Hedges & Becker, 1986; Johnson, 1989).

For studies utilizing a factorial design, the entire ANOVA table should be reported, or all cell means and either (a) the within-cell standard deviations, (B) the standard deviation for each level of the moderator independent variable, or (c) the MS_e for the ANOVA. This is necessary in order to decompose the design allowing the researcher either to aggregate the effect of the independent variable over the other variables (main effects) or “represent the effect of interest within levels of the other variables (i.e. as simple main effects)” (Johnson, 1989, p. 17). The study by Price (1985) failed to report the ANOVA table and as a result could not be included in the analysis.

During the analysis, two other studies were rejected. One study Silins (1994 a) was based on the same sample, addressed the same research issue, and used the same analytical procedures as her other study reported in another journal. The second study, a microfiche of a dissertation, was so poorly reproduced that a guessing factor would have had to be included in order to calculate the effect sizes.

Requests for correlation matrices were sent to two researchers. One researcher responded apologizing that the matrix was no longer available (Appendix C). There was no response from the other. Consequently the review included 38 published journal articles, dissertation studies, and papers presented at peer-reviewed conferences.

Coding Study Characteristics

In order to determine the relationship among the results of studies the characteristics of the studies have to be coded. For this study, the approach used by Hallinger and Heck (1990), based on the framework developed by Pitner (1988), was adopted (Appendix A). In this approach, the studies were classified according to variations on three models. Model A, direct effects: the principal's actions affect student outcomes. Model B, mediated effects: the principal's actions affect on student outcomes mediated through other variables. Model C, reciprocal effects: the principal's actions influence teacher behavior which then affects the principal's actions, and through these processes, outcomes are affected. Variations of the model emerge with the inclusion of antecedent effects on the principal's behavior.

This approach allowed for the observation of the changing conceptualization and the varying interpretations applied to the functions of leadership over the 15 year period, 1985-2000. In addition to this, it allowed for an analysis of the relationship between the strength of the effect and the design of the study.

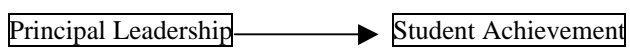
The Models

The direct effects model (Figure 6, Model A) makes the assumption that the actions of leaders have a direct effect on student outcomes. As articulated by Hallinger and Heck (1998), there is also the assumption that “these effects can be measured reliably apart from other related variables” (p. 163). This, of course, is a simplification of the influence of leadership on the overall effectiveness of the school, and fails to recognize the fact that other variables may contribute to student outcomes.

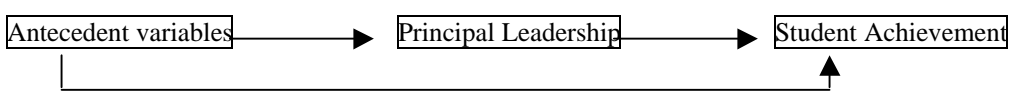
The mediated- effects model (Model B, Figure 6) hypothesizes that the effect of the school-site leader is indirect, being mediated by other school-level variables such as teacher commitment, school organizational culture, and instructional practices (Leithwood, 1994; Sheppard, 1996). This view of the effects of leadership recognizes Glickman’s (1991) contention that the role of the principal is that of a coordinator and facilitator. The emphasis is in infusing the entire organization with a zealotry to succeed. This model is not only stronger in design, but is also a realistic conceptualization of the manner in which leadership functions.

The reciprocal-effects model proposes that the relationship between the principal and the intervening variables is interactive. This model suggests that as the leader’s influence changes the organization taking it to another level, the organization then influences the leader to act in new ways. The assumption, therefore, is that leadership is an adaptive process and the causal relationships are multi-directional. As stated by Hallinger and Heck (1998) “further development of this model represents a profitable area for future research” (p. 168).

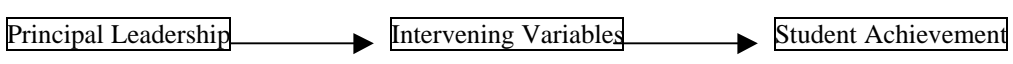
Model A: Direct-effects Model



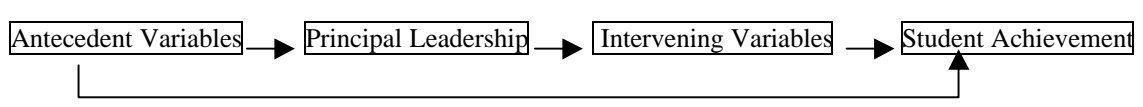
Model A-1: Direct-effects with Antecedent effects



Model B: Mediated-effects



Model B1: Mediated-effects with Antecedents



Model C: Reciprocal-effects Model

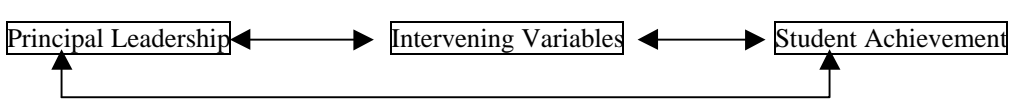


Figure 6: Modeling principals effects on school effectiveness (adapted from Pitner, 1998, pp. 105-108). From P.Hallinger and R. H. Heck, 1998, "Exploring the principal's contribution to school effectiveness": 1980-1995. *School Effectiveness and School Improvement*, 9, (2), 164-165.

In addition to the conceptual framework for viewing the effects of the principal on student outcomes, the studies were also categorized and coded based on substantive and methodological features. Substantive features that were coded included: publication features (year, journal article, presentation, dissertation), sample type, school type (elementary, middle or secondary), and dependent variable (i.e. overall school effectiveness or specific measures of school effectiveness).

Academic performance on measures such as subject-based tests or normed assessment batteries was classified as achievement. Where the measure of effectiveness was related to other school outcome variables, for example, engagement in school, and identification with school, this was coded as affect. Attendance was coded as attendance and where a study looked at number of outcomes, these were coded other or various outcomes. Methodological features included: instrumentation, sample size, number of observations aggregated into each data point in the statistical analysis, and analytic technique.

In addition to the above, for the purposes of calculating the effect sizes, leadership was classified based on the dimensions as reflected by the measurement instruments used by the various researchers. For each study the effect size was calculated. Also, where dimensions measured were common across instruments, the data were aggregated and a mean effect size calculated and reported. However, because the 38 studies used approximately 16 instruments as direct measures of leadership, it would have served no useful purpose in calculating the mean effect size across instruments. Therefore this step, although important in determining whether the instrument influenced the magnitude of the effect of leadership on the outcome variable, was not done.

It was possible to adopt the above classification framework because of the commonality between and among the different conceptual approaches to leadership. To reiterate Leithwood et al. (1999), the various leadership models, although representing to varying degrees the differences in the conceptualization of leadership, a consequence of which is an emphasis on different behavioral attributes of the leader, include dimensions that are common to all.

Concurring with the advice of Hunter and Schmidt (1990), issues pertaining to the reliability of the measures and validity of the studies were also noted. Also, being cognizant of the “garbage in -- garbage out” criticism of meta-analysis, only studies that met all the criteria were included, and additionally, based on the coded characteristics, the studies were tested during the analysis to detect any possible bias.

Calculating the effect size

Over (67.5%) of the studies were correlational, 10% were based on means and standard deviations, and the remaining 22.5% reported either the t or F statistics. The effect size calculated was the g, the estimate of the estimate of the population standardized mean difference: in this case, the difference between the student outcome means (based on the leadership behaviors), divided by the pooled standard deviation (Hedges & Olkin, 1985). The sign of the difference was positive when certain leadership behaviors had a positive effect on student outcomes.

The computation of g was based on (a) \bar{E} and \bar{t} for those studies that estimated mean differences, (b) means and standard deviations for the descriptive studies, and (c) the conversion of r for the correlational studies. For studies that manipulated more than one approach to, or style of leadership, or addressed a number of outcomes, for example Blank (1987), a separate effect size was computed for each manipulation.

Whenever possible, the pooled standard deviation was estimated from the portion of each study's data that entered an effect size. Consequently, if the effect of an aspect of leadership was calculated within the context of the elementary schools, the pooled standard deviation was estimated from the standard deviations for the elementary schools if the information was available.

As explained by Hedges and Olkin (1985), g is a slightly biased estimator of δ ; the population standardized mean difference especially for small samples for which it tends to overestimate. To correct for this bias, all g s were converted to d s by multiplying g by the constant c_n . Therefore, $d = c_n g$

$$c_n = 1 - \frac{3}{4_{n^a} + 4_{n^b} - 9}$$

where n^a and n^b are the sample size of the two groups (Hedges & Olkin, 1985).

To determine the homogeneity of each set of d s, the Q_w statistic was calculated. This statistic examines whether the studies shared a common effect size and has a chi-square distribution with $k-1$ degrees of freedom, where k is the number of effect sizes (Hedges & Olkin, 1985). The chi-square statistic is sensitive to k , therefore a large number of effect sizes combined into a category is more likely to result in significant homogeneity statistics.

In the absence of homogeneity, the procedures outlined by Hedges and Olkin (1985) were used in testing the models. Therefore, the study characteristics were examined in relation to the variability of the heterogeneous effect sizes with a view to determining the relationship if any, between the study characteristics and the magnitude of the effect sizes (Hedges & Olkin).

For categorical models, the Q_B statistic was calculated. This statistic has an approximate chi-square distribution of $p-1$ degrees of freedom, where p is the number of classes, and, analogous to the main effects in an ANOVA, estimates the between-classes effect. A significant Q_B statistic is an indication that the effect sizes are heterogeneous, and consequently would have to be tested for homogeneity within the classes or subgroups defined by the study characteristics. In reporting tests of categorical models, the mean weighted effect size for each class is also calculated, together with the 95% confidence interval (CI) for each mean. The non-inclusion of zero in the CI indicates that the mean weighted effect size is significantly different from zero.

Transformation to d

With reference to the definition of effect size as the standardized mean difference, Hattie (1998) reminds that the preferred method of computing the effect size is to use the means and standard deviations of an outcome variable. In this regard, he has compiled a list of formulae for the transformation of various statistics to d , the statistic of interest. For studies that reported only means and standard deviations the formula used was

$$d = \frac{\overline{X}_1 - \overline{X}_2}{S_{pooled}}$$

where X_1 is the mean of the first group (the treatment group), X_2 is the mean of the second group (the comparison group), and S_{pooled} is the pooled standard deviation for the two groups. For two studies (Freeman, 1987; Heck, 1992), it was necessary to calculate the pooled standard deviation. Thus the following formula was applied to the given descriptive statistics.

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)(s_1)^2 + (n_2 - 1)(s_2)^2}{n_1 + n_2 - 2}}$$

where n_1 is the sample size for the first group

n_2 is the sample size for the second group

s_1 is the standard deviation for the first group

s_2 is the standard deviation for the second group

For studies that reported inferential statistics the following formulas were used. For those that reported t values, the formula used was

$$d = t \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

where t is the t -value between the two groups

n_1 is the sample size of the first group

n_2 is the sample size of the second group

Those studies that utilized the one-way ANOVA procedure in a two-group comparison, the formula for the computation of the effect size was the same as that of the t -test but with the substitution of the square root of the F -value for the t -value. Thus the following was the formula for the calculation of d

$$d = \sqrt{\frac{F(n_1 + n_2)}{n_1 n_2}}$$

For ANOVAs that compared three or more groups, the means for each group and the means-square within groups (MS_w) was applied to the formula

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{MS_w}}$$

where X_1 and X_2 are the means of the comparison groups of interest.

For one such ANOVA, (Blank,1987), the MS_w had to be computed from the group means and the F -value. This was done by estimating the sums-of-squares between groups (SS_b), and computing the mean squares by dividing the SS_b by $k-1$ and then dividing the MS_b by F .

$$SS_b = \sum (M_k - GM)^2 n_k$$

where M_k is the mean of each group, GM is the grand mean, and n_k is the sample size of each group. To compute the MS_b , the following formula was applied.

$$MS_{between} = \frac{SS_{between}}{k - 1}$$

where k is the number of groups. It is now possible to compute the MS_w

Many studies when examining the influence of leadership on student outcomes, deconstructed leadership into various dimensions and applied the multiple regression analytical procedure to test the relationship to the criterion. However, the results of a multiple regression cannot be converted to an effect size unless the zero-order correlation between the independent variable and the dependent variable is reported, or the beta-weights (bs) of the bivariate

regressions, which are equivalent to r -values, are given (Johnson, 1989). Therefore for correlational studies, the following formula was used.

$$d = \frac{2r}{1 - (.27)^2}$$

For two studies, Garner (1989) and Philbin (1997), the effect sizes had to be calculated from the reported N s and p -values. Although this method does not give the most reliable estimates, in the absence of the alternative statistical information, adopting the procedure is acceptable. This procedure is complicated; the p -value is first transformed into z , the z is then transformed into r , and from the r the g is calculated. Consequently, the transformation was done using the DSTAT software for the meta-analysis (Johnson, 1989).

There were no studies that utilized the factorial ANOVA procedure or reported only frequencies or proportions. Therefore these formulae are not presented in this overview of the methodology.

Furthering the Analysis

Meta-analysis limits the researcher to the analysis of studies for which the d or r statistic can be computed. However, as researchers continue to reconceptualize and redefine leadership, the analytical procedures adopted have become more sophisticated. While this development was inevitable and is correct, in the absence of the reporting of the requisite statistical information for the conduct of the meta-analysis, the studies cannot be utilized in the synthesis. While this is so, these studies, instead of being rejected, were coded, the analytic techniques noted, and the effects reported. This approach is similar to that adopted by Hallinger and Heck (1998), and Leithwood et al. (1990). Thus in the words of Leithwood and his colleagues, “a form of disciplined “stock-taking was the primary method used” (p.6) in the synthesis of these studies.

The Synthesis

After encoding the data and aggregating these according to the eight dimensions discussed previously, the mean effect sizes, variances, and confidence intervals were then computed. In order to avoid the violation of the assumption of independence that could accrue from using the many d s aggregated into one construct, Hedges (1985) suggests either randomly selecting an effect size from each study or using the median effect. In this study, the overall mean effect size is based on the median effect size from each study. Therefore, presented in the results section are the weighted means, the variance, the confidence interval and the overall mean effect. The mean effect was first calculated using the formula

$$\bar{d} = \frac{\sum[N_i d_i]}{\sum N_i}$$

where d_i is the corrected median Hedges' d , and then computed using DSTAT. All further calculations were computed using the software.

Analysis of the Results

In analyzing the results of each aggregation, a comparison is made of the corrected deviation to the mean to assess the size of the potential variation across studies. "If the mean is more than two standard deviations larger than zero, then it is reasonable to conclude that the relationship considered is always positive" (Hunter & Schmidt, p. 82). However, if the distribution of scores approaches that of the normal curve, then a mean one standard deviation above zero is also considered to be important. Analysis of the data was done via (a) hand calculations, (b) utilization of the Statistical Package for the Social Sciences (SPSS 10.0), and (c) Software for the Meta-analytic Review of Research Literature: DSTAT (Johnson, 1989).

Summary

This chapter reflected to a large extent the methodology as proposed by Hedges and Olkins (1985), Hedges and Becker (1986), Johnson (1989) with references from Hunter and Schmidt (1990). The presentation of studies that did not meet the criteria of the meta-analysis reflected the approach adopted by Hallinger and Heck (1998), and Leithwood, Jantzi, and Steinbach (1999).