ISSN No.
0146-9283

Fall
2011

educational considerations

published at kansas state university college of education

educational considerations

Vol. XXXIX, Number 1, Fall 2011

Available online at:
http://coe.ksu.edu/EdConsiderations/

TABLE OF CONTENTS

Special Issue on Class Size and Student Achievement
by
James L. Phelps

Educational Considerations Design/Layout by
Mary Hammel, Kansas State University

Educational Considerations invites subscribers for only **$13.00** annually.
Educational Considerations is published and funded by the College of Education at Kansas State University. Address correspondence to Editor, *Educational Considerations*, Bluemont Hall, Kansas State University, Manhattan, KS 66506 or call (785) 532-5543.

PUBLICATION INFORMATION

Educational Considerations is a peer-reviewed journal published at the College of Education, Kansas State University. ***Educational Considerations*** and Kansas State University do not accept responsibility for the views expressed in articles, reviews, and other contributions appearing in this publication. In keeping with the professional educational concept that responsible free expression can promote learning and encourage awareness of truth, contributors are invited to submit conclusions and opinions concerned with varying points of view in and about education.

Educational Considerations is published two times yearly. Editorial offices are located at the College of Education, Bluemont Hall, 1100 Mid-Campus Drive, Kansas State University, Manhattan, KS 66506-5301. Correspondence regarding manuscripts should be directed to the Executive Editor at fecrampton@gmail.com. No remuneration is offered for accepted articles or other materials submitted.

By submitting to *Educational Considerations*, the author guarantees that the manuscript has not been previously published. The University of Chicago's *Manual of Style*, 16th edition is the editorial style required. Authors may select from two citation systems: note (footnote) or author-date, as described in Chapters 14 and 15 of the manual, titled "Documentation I" and "Documentation II," respectively. For note style, footnotes with full details of the citation should be listed at the end of the manuscript. No bibliography is needed. Tables, graphs, and figures should be placed in a separate file. An abstract of 150 words must accompany the manuscript. **Manuscripts should be submitted electronically to Faith Crampton at fecrampton@gmail.com as an e-mail attachment.** Complete name, address, telephone number, and email address of each author should be included in the body of the e-mail and on the title page of the manuscript. Photographs, drawings, cartoons, and other illustrations are welcome. Authors are required to provide copies of permission to quote copyrighted materials. Queries concerning proposed articles or reviews are welcome. The editors reserve the right to make grammatical corrections and minor changes in article texts to improve clarity. Address questions regarding specific styles to the Executive Editor.

Subscription to ***Educational Considerations*** is $13.00 per year, with single copies $10.00 each. Correspondence about subscriptions should be addressed to the Business Manager, c/o The Editor, *Educational Considerations*, College of Education, Kansas State University, Manhattan, KS 66506-5301. Checks for subscriptions should be made out to ***Educational Considerations***.

Printed in the United States of America.

Introduction to the Special Issue

Faith E. Crampton, *Executive Editor*
David C. Thompson, *Chair, Board of Editors*

We are pleased to share with you this special issue revisiting the research on the relationship between class size and student achievement, along with its implications for education policymakers and practitioners. For over half a century, researchers have struggled to identify those variables that contribute in significant ways to students' academic success, and the resulting, voluminous literature is rife with contradictory results. At the same time, the positive results of class size research, which is part of the body of "production function" analysis, has received broad acceptance by policymakers, parents, and practitioners who believe "smaller is better."

The fiscal implications of this belief for state and local school districts have been enormous. As such, the re-examination of class size research is particularly relevant at a time when many states and localities are making significant cuts in education budgets that require hard choices as to which programs and initiatives can be reduced or eliminated without harming students. As states, schools, and local districts make these difficult decisions, it is essential that they balance cost-effectiveness with the best interests of students and maintain the ethical, moral, and legal imperatives of equality of educational opportunity and social justice.

To that end, this issue contains five interwoven articles by James L. Phelps, whose distinguished educational career has included serving as Special Assistant to Governor William Milliken of Michigan and Deputy Superintendent in the Michigan Department of Education. This special issue of *Educational Considerations* is unique in the sense that rather than a collection of articles, it more closely resembles a monograph comprised of five chapters. Dr. Phelps' perspective, which melds research, practice, and policy, is also unique, making his analysis of the past, present, and future of class size reduction research and initiatives invaluable.

The special issue opens with an article titled, "Another Look at the Glass and Smith Study on Class Size." Glass and Smith's iconic 1978 study[1] set the stage for much of the narrative around the impact of class size on student achievement which was later reinforced by results from the Tennessee STAR experiment.[2] The second article, "A Practical Method of Policy Analysis by Considering Productivity-Related Research," presents a fresh approach to the type of analysis that historically has underpinned much of class size research. In the third article, "A Practical Method of Policy Analysis by Estimating Effect Size," Phelps takes a critical look at the use of "effect size," an oft-used metric in class size research to judge its success in raising student achievement, and offers alternative methods for calculating and interpreting it. The fourth article, "A Practical Method of Policy Analysis by Simulating Policy Options," provides an example of how the cost-effectiveness of education reforms like class size reduction can be simulated statistically in ways that are robust and meaningful to education decision makers.

The final piece is a closing essay that summarizes the findings of the previous articles and reinforces the importance of the development of a unified theory of the production of student achievement, a thread that runs through all of the articles. At the same time, Phelps acknowledges the difficulty involved in operationalizing such a theory through research methods and statistical analyses that capture the complexity of human endeavors, making the research on class size and student achievement an ongoing endeavor.

Endnotes

[1] Gene V. Glass and Mary Lee Smith, *Meta-analysis of Research on the Relationship of Class-size and Achievement* (San Francisco, CA: Far West Laboratory for Educational Research and Development, 1978).

[2] C.M. Achilles, B.A. Nye, J.B. Zaharias, and B.D. Fulton, "The Lasting Benefits Study (LBS) in Grades 4 and 5 (1990–1991): A Legacy from Tennessee's Four-year (K–3) Class-size Study (1985–1989)," Project STAR, a paper presented at the North Carolina Association for Research in Education, Greensboro, North Carolina, January 14, 1993.

Another Look at the Glass and Smith Study on Class Size

James L. Phelps

One of the most influential studies affecting educational policy is Glass and Smith's 1978 study, *Meta-Analysis of Research on the Relationship of Class-Size and Achievement*.[1] Since its publication, educational policymakers have referenced it frequently as the justification for reducing class size. While teachers and the public had long believed lowering class size would be advantageous, Glass and Smith gave the idea legitimacy. This article is a review and reanalysis of the Glass and Smith study. While this review maybe considered much too late, it does serve the purpose of re-evaluating a frequently cited study to either support or challenge various aspects of the original findings. To that end, the article is divided into six major parts. It begins with an overview of the Glass and Smith study for those who may not be familiar with the specifics. This is followed by a description of their findings and comments upon these by the author. The fifth section presents a reanalysis of their data. The article closes with observations and conclusions.

Overview

To capture the character of the original study, the summary from Glass and Smith is presented here in its entirety (pp. iv-vi):

> Research on the relationship between class-size and academic achievement is old, huge and widely believed to be inconclusive. Previous reviews of the evidence have been overly selective and insufficiently quantitative. Timid qualifications were offered where bold generalizations were possible. In the summer of 1978, the *New York Times* gave front-page coverage to a study published by Educational Research Services, Inc. (Porwell, 1978). This organization is funded jointly by the American Association of School Administrators, the Council of Chief State School Officers, and several other professional administration groups. The "Porwell Report" staggered visibly under the weight of the research data and eventually arrived at the following conclusion sad for teachers to behold:

James L. Phelps holds a Ph.D. from the University of Michigan in Educational Administration. He served as Special Assistant to Governor William Milliken of Michigan and Deputy Superintendent in the Michigan Department of Education. Active in the American Education Finance Association, he served on the Board of Directors and as President. Since retirement, he spends a great deal of time devoted to music, composing and arranging, playing string bass in orchestras and chamber groups, as well as singing in two choirs. He resides with his wife, Julie, in East Lansing, Michigan.

(Quotation, continued)

> Research findings on class size to this point document repeatedly that the relationship between pupil achievement and class size is highly complex.
>
> There is general consensus that the research findings on the effects of class size on pupil achievement across all grades are contradictory and inconclusive.
>
> Existing research findings do not support the contention that smaller classes will of themselves result in greater academic achievement gains for pupils (Porwell 1978, 68-69).

The research reported herein contradicts the conclusions of the Porwell Report. Indeed, it establishes clearly that reduced class-size can be expected to produce increased academic achievement. In pursuing this conclusion, we discovered many of the reasons why previous research reviewers lost their way in the forest of data and failed to find a defensible generalization.

We collected nearly 80 studies on the relationship between class-size and achievement. These studies yielded over 700 comparisons of the achievement of smaller and larger classes; these comparisons rest on data accumulated from nearly 900,000 pupils of all ages and aptitudes studying in all manner of school subject. Using complex methods of regression analysis, the 700 comparisons were integrated into a single curve showing the relationship between class-size and achievement in general. This curve revealed a definite inverse relationship between class-size and pupil learning. Similar curves were derived for a variety of circumstances hypothesized to alter the relationship between achievement and class-size. Virtually none of the special circumstances altered the basic relationship; not grade level, nor subject taught, nor ability of pupils. Only one factor substantially affected the curve, viz., whether the original study controlled adequately (in the experimental sense) for initial differences among pupils and teachers in smaller and larger classes. The nearly 100 comparisons of achievement from the well-controlled studies thus form the basis of our conclusion about how class-size is related to academic achievement. This curve appears in the Figure below. As class-size increases, achievement decreases. A pupil, who would score at about the 83rd percentile on a national test when taught individually, would score at about the 50th percentile when taught in a class of 40 pupils. The difference in being taught in a class of 20 versus a class of 40 is an advantage of 6 percentile ranks. The major benefits from reduced class-size are obtained as size is reduced below 20 pupils.

As one looks at the representation of the relationship between achievement and class size, several immediate questions arise:

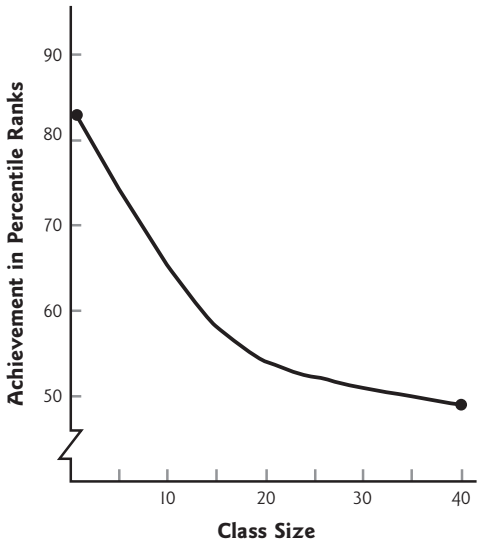
(1) Why are the relationships all above the 50th percentile?

(2) Why is the relationship curved?

(3) Why are the relationships not reported for class sizes larger than 40?

(4) How many teachers are necessary to bring the class size down from 40 to 20, from 40 to 10, and from 40 to 1?

Figure 1
Curve Derived by Glass and Smith from 100 Comparisons from Well Controlled Studies



Source: Gene V. Glass and Mary Lee Smith, *Meta-Analysis of Research on the Relationship of Class-Size and Achievement* (San Francisco, CA: Far West Laboratory for Educational Research and Development, 1978), vi, Figure 1.

Research Method

Glass and Smith described their research method, meta-analysis, in detail.[2] They took comparisons between achievement and class size from many studies, formed a new data set, and then conducted a regression analysis using this data set. The following subsections summarize each of the topics addressed.

Defining the class size field (p. 9). Glass and Smith selected the number of pupils within a class with one teacher as the measure rather than a measure of "staff adequacy," the number of teachers per 100 pupils. While there was a mathematical transformation equating the two notions, there was a substantial difference in their policy implications, to be discussed later.

Coding characteristics of studies (pp. 10-13). Glass and Smith collected data for the following fields, although data from some studies were not available and not all fields were completely filled: ID number of study; year of study (1900-1979); source of data (whether from journal, book, thesis, or unpublished source); subject taught (reading, mathematics, language, psychology, natural/physical science, social science and history, and "all others"); duration of instruction, in hours and in weeks; number of pupils, instructional groups, and teachers; pupil/instructor ratios for small and large classes; assignment of pupils and teachers; subject of achievement measure; and achievement measure (the difference in achievement between the small and large classes). Other data items were collected but are not included in this listing because they were not incorporated into their analyses.

Quantifying outcomes (pp. 13-14). For each of the comparisons from each of the studies a single statistic was required. Glass and Smith stated:

> No matter how many class-sizes are compared, the data can be reduced to some number of paired comparisons, a smaller class against a larger class... The most obvious differences involve the actual sizes of the "smaller" and "larger" classes and the scaled properties of the achievement measure... The measurement scale properties can be handled by standardizing all mean differences in achievement by dividing by the within group standard deviation (a method that is complete and discards no information at all under the assumption of normal distributions).

The achievement measure was standardized across all studies through the use of standard or Z-scores. The achievement measure in Z-scores was notated by Glass and Smith (pp. 13-14) as:

$$\Delta_{(S-L)} = (\overline{X}_{(S)} - \overline{X}_{(L)} / \hat{\sigma}$$

where

S represents the small class;

L represents the large class;

$\overline{X}$ represents the achievement mean;

and $\hat{\sigma}$ represents the standard deviation.

Calculating the achievement measure $\Delta_{(S-L)}$. Because many of the studies from which the data were taken did not include basic descriptive statistics, alternative methods to calculate the achievement variable had to be developed. Glass and Smith described their methods on pages 14-15.

Describing the class size and achievement relationship. Glass and Smith considered several alternative statistical techniques to describe the aggregated findings. The selected alternative is quoted below (pp. 15-19):

> Finally, regression equations could be constructed in which $\Delta_{(S-L)}$ is partitioned into a weighted linear combination of S and L and function thereof and error... But the regression of $\Delta_{(S-L)}$ into only S and L requires three dimensions to be depicted. Anything more complex than a simple two-dimensional curve relating achievement to the size of class was considered undesirably complicated and beyond the easy reach of most audiences who hold a stake in the results.
>
> The desire to depict the aggregate relationship as a single line curve is confounded with the problem of essential inconsistencies in the design and results of the various studies. A single study of class-size and achievement may yield several values of $\Delta_{(S-L)}$... This set of Δ's from a single study will form a consistent set of values in that they can be joined to form a single connected graph depicting the curve of achievements as a function of class-size. However, various values of $\Delta_{(S-L)}$ arising from difference studies can show confusing inconsistencies. For example, suppose that Study #1 gave Δ (10-15), Δ (10-20), and Δ (15-20) and Study #2 gave Δ (15-30), Δ (15-40), and Δ (30-40). A few moments reflection will reveal that *there is no obvious or simple way to connect these values into a single connected curve* [emphasis added].[3]
>
> The eventual solution to these problems proceeded as follows: $\Delta_{(S-L)}$ was regressed onto a quadratic function of S and L by means of the least-squares criterion: then that set of values of $\hat{\Delta}$ that could be expressed as a single, connected curve was found.

The regression model selected accounted for variations in $\Delta_{(S-L)}$ by means of S, S^2 and L. *Obviously, something more than a simple linear function of S and L was needed, otherwise a unit increase in class-size would have a constant effect regardless of the starting class-size S; and the S^2 term seemed as capable of filling the need as any other. The size differential between the larger and smaller class, L-S was used in place of L for convenience* [emphasis added]. Thus, the $\Delta_{(S-L)}$ values were used to fit the following model: [4]

$$\Delta_{(S-L)} = \beta 0 + \beta_1 S + \beta_2 S^2 + \beta_3 (L-S) + \varepsilon... \quad (1)$$

The problem now is to find the set of $\hat{\Delta}$'s in this surface that can be depicted as a single curved-line relationship in a plane.

It is important at this point to determine the dimensions of the equation. Obviously, achievement is the first dimension. Class size (the S and S^2 terms forming a parabola) is the second because for any value of S a value for achievement can be calculated. The uncertainty pertains to a possible third dimension. L would be a third dimension if it were a data variable entered into the regression equation and a value for achievement could be calculated for each value of L. However, L was not a data variable entered into the regression; rather, (L-S) was the variable. This point is critical: (L-S) can produce a value for achievement if, and only if, L is fixed and S varies. Therefore, (L-S) is not an independent third dimension; instead, it is a line within the class size dimension.

Next, Glass and Smith described a "consistency property." The relevant section of their study (pp. 17-19) has been included here because of its importance to the commentary in the fourth section of this article:

> The property that must hold for a set of $\hat{\Delta}$'s before they can be depicted as a connected graph in a plane is what might be called the consistency property [emphasis in the original]:
>
> $$\Delta_{n1-n2} + \Delta_{n2-n3} = \Delta_{n1-n3}$$
>
> for n1 < n2 < n3. If this property is not satisfied, then one is in the strange situation of claiming that the differential achievement between class size 10 and 20 is not the sum of the differential achievement from 10 to 15 and then from 15 to 20.
>
> When the consistency property is imposed on the regression equation, it follows that:
>
> $$\hat{\beta}_0 + \hat{\beta}_1 n_1 + \hat{\beta}_2 n_1^2 + \hat{\beta}_3 (n_2 - n_1) + \hat{\beta}_0 + \hat{\beta}_1 n_2 + \hat{\beta}_2 n2^2 + \hat{\beta}_3 (n_3 - n_2) =$$
> $$\hat{\beta}_0 + \hat{\beta}_1 n_1 + \hat{\beta}_2 n_1^2 + \hat{\beta}_3 (n_3 - n_1) \quad (3)$$
>
> Simple algebraic reduction produces the following:
>
> $$\hat{\beta}_0 + \hat{\beta}_1 n_2 + \hat{\beta}_2 n_2^2 = 0 \quad (4)$$
>
> The two solutions to the quadratic equation...are points n_2 such that if $\hat{\Delta}_{(S-L)}$ is measured with n_2 as either the larger L, or smaller, S, class-size then the resulting set of $\hat{\Delta}$'s will lie on the four dimensional regression curve...but can be depicted as a single line curve in a plane. Since n_2 becomes the point around which values of n_1 and n_3 are selected, it will be called the pivot point [emphasis in the original]. That there are two solutions for n_2 is perplexing; fortunately in the analyses to be reported the two corresponding curves were virtually parallel in practice.
>
> A single line curve in a plane can be constructed by solving for one or the other values of n_2 in (4) and constructing a set of Δ values. These values will give the standardized mean differences in achievement between n_2 and any other class-size. The curve that connects these Δs has no non-arbitrary starting point. One can assume for convenience sake that the achievement curve (z), instead of the differential achievement curve (Δ) is centered around an arbitrary class-size, e.g., something like the national average in the low 20s. Finally, for descriptive purposes, the metric of percentile ranks was chosen over the metric of z-scores; thus, the curve z was transformed into a curve of percentile ranks by assuming a normal distribution of achievement.[5]
>
> Comment on Statistical Inference [Underline in original]
>
> In the analyses that follow, ordinary matters of statistical inference have been ignored. The application of usual interval estimation procedures or statistical tests makes little sense for two reasons. The data base is laced with a complicated structure of interdependent observations; several comparisons arise from a single study when more than two class-sizes are compared, and there is no sensible way to reduce each study to one observation... Secondly, randomization is absent from the data set in any form that would make probabilistic models based on it applicable.

Findings

According to Glass and Smith (p. 20), "The report of findings falls into two broad categories: (1) description of the data base and (2) regression analyses relating to achievement and class-size." I begin here with a quotation from their description of the data base (p. 20):

> In all, 77 different studies were read, coded, and analyzed. These studies yielded a total of 725 Δ's. The comparisons are based on data from a total of nearly 900,000 pupils spanning 70 years research in more than a dozen countries. (The entire set of data is reproduced in the appendix to this report.)

Table 1
Glass and Smith Regression Equation Results

Class Size	*Delta*	*Interval*	*Difference*
1	0.5859	1 to 65.81	0.00001
10	0.2895	1 to 10	0.2964
20	0.0723	10 to 20	0.2172
25.84	0.0000	20 to 25.84	0.0723
30	-0.0269	20 to 30	0.0269
33.41	-0.0338	30 to 33.41	0.0068
40	-0.0081	30 to 40	-0.0256
40.97	0.0000	40 to 40.97	-0.0081
50	0.1287	40 to 50	-0.1287
60	0.3835	50 to 60	-0.2548
65.81	0.5857	60 to 65.81	-0.2022
Sum			0.0003

Source: Glass and Smith (1978).

Several tables were presented in the study showing the frequency distributions of the data characteristics (Tables 1-5, pp. 20-26). These are not summarized here. However, in the data set, small class size ranged between 1 and 70. Large class-size ranged between 2 and 146. These values come into consideration when parameters are set in the regression equations.

In their regression analyses section (p. 29), Glass and Smith presented the statistical properties of the dependent variable $\Delta_{(S-L)}$. Most interesting, 40% of the values for $\Delta_{(S-L)}$ were negative, and 60% were positive. The large percentage of negative values for Δ raises an interesting situation. For any value of S, if the sum of the Δ's is positive, the slope of the line will be positive; however, if the sum of the Δ's is negative, the slope of the line will be negative. This circumstance raises the possibility that the curve representing the full range of class sizes will be comprised of both positive and negative slopes.

> The result of the regression analysis for the entire data set was (p. 33):
>
> $\Delta_{(S-L)} = .57072 - .03860\ S + .00059\ S^2 + .00082\ (L-S)$

At this point, Glass and Smith provided a table with a range of small and large class-sizes with the Δ as calculated from the regression results above (p. 34). The small class size (S) is only up to 30, and the large class size is (L) up to 40, even though these values are substantially higher in the data set. This table, in an expanded form, is provided below. (See Table 1.) In order to calculate the regression results, a value for the large class size must be set, in this case a class size of 65.81, for a reason to become clear later. Calculations have also been included to test the consistency property: If intervals A to B + B to C = A to C.

> Glass and Smith concluded:
>
> > These data show that the difference in achievement between class-size 1... and class-size of 40 is more than one-half standard deviation. The difference between class-size 20 and 40 is only about five hundredths standard deviation. Class-size differences at the low end of the scale have quite important effects on achievement; differences at the high end have little effect (p. 34).

It should be noted in Table 1 that the predicted achievement for a class size of 40 is marginally better than that for a class-size of 30; and achievement continues to increase to a class-size 65.81 where achievement is virtually the same as for a class-size of 1.

Most interestingly, when the consistency property is tested using the data from the table, the sum of the intervals of class size from 1 to 10 and 10 to 20 equals the interval from 1 to 20, and all other intervals as well.[6] As will be demonstrated later, the data from Table 1 can be graphed in two dimensions.

Three questions arise: (1) Why does the regression equation predict almost the same achievement level for a class-size of 1 and 65.81; (2) Why are there two predicted achievement values of 0; and (3) What is the consequence of setting the value of L?

Utilizing the consistency property, Glass and Smith (p. 34) observed: "The curved regression surface can be reduced to a single line curve in a plane by imposing the consistency condition and solving for the pivot points. The two pivot points are the solutions to .57072 - .03860 (P) + .00059 (P^2) = 0." They calculated the pivot points to be approximately 43 and 23. Because a parabola was selected as the curve for the regression analysis, it comes as no surprise that the curve on its downward path intersected the zero plane of the Z-axis, continued downward to a minimum point, about 33.4, and then moved upward, again intersecting the zero plane of the Z-axis as it continued upward.[7] The pivot points are the intersections of the parabola with the Z-axis. As part of the results of their study, Glass and Smith (p. 35) presented a table showing the results of the consistency property transformation, although no calculations were presented. This statement preceded and followed the table, which has been expanded here as Table 2 to show class-sizes larger than 40:

Table 2
Glass and Smith Results Including the Consistency Property Transformation

Small Class Size	*Large Class Size*	*Standardized Differential Achievement, Δ_{S-L}*
1	23	0.551
10	23	0.254
20	23	0.037
23	23	-0.005
23	30	0.001
23	40	0.009
23	50	0.017
23	60	0.025
23	65.81	0.030

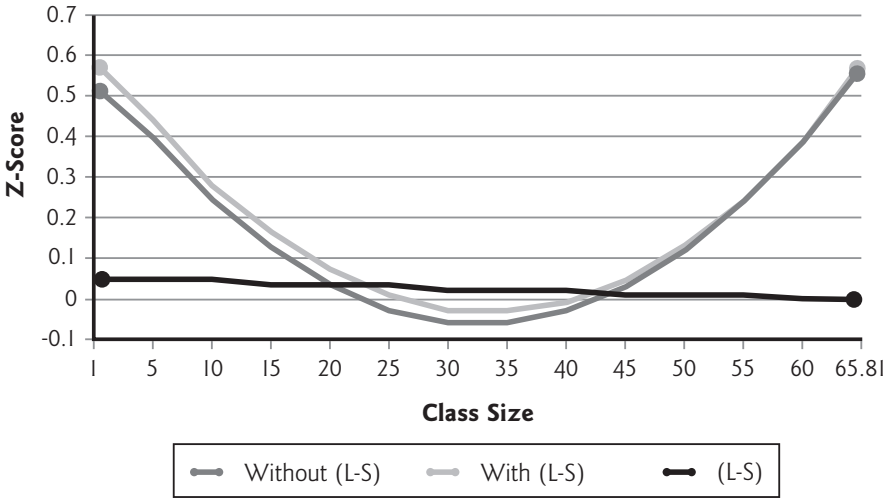
Source: Glass and Smith (1978, 35).

> The lower value, 23 was selected as the pivot point around which to construct the connected curve; the choice was arbitrary and calculations not reported here revealed it to be largely immaterial. The values are for $\Delta_{(S-P)}$ and $\Delta_{(P-L)}$ are as follows for P = 23:
>
> $\Delta_{1-23} = .551$
> $\Delta_{2-23} = .513$
> $\Delta_{5-23} = .407$
> $\Delta_{10-23} = .254$
> $\Delta_{20-23} = .037$
> $\Delta_{23-30} = .001$
> $\Delta_{23-40} = .009$
>
> Hence, on this curve the difference between achievement in class-sizes 1 and 40 is .551 + .009 = .560... The ordinate is represented by a standard score metric; the zero point (of the graph) is arbitrarily fixed at a class-size of 30 (p. 35).

The reader is urged to pay particular attention to the shift in the calculations due to introduction of the condition: $\Delta_{(S-P)}$ and $\Delta_{(P-L)}$ where P = 23. In the first case, P is substituted for L, and, in the second, P is substituted for S. Therefore, up to S = 23, the variable S changes, and L is fixed; above 23, S is fixed, and L changes. Below S = 23, the relationship is curved (parabolic) while above S= 23 the relationship is linear. In essence, at S = 23, the regression equation changes.

Glass and Smith presented finding for subsets of the data, including "elementary vs. secondary grades" and "well-controlled studies vs. poorly-controlled studies" (pp. 38-42). Several graphs were pre-

Figure 2
Glass and Smith Regression Equation



sented to support their findings, based on the consistency property transformation, not on the derived regression equations. The graphs depict predicted achievement in terms of Z-scores. Finally, because the Z-axis was measured in Z-scores, the final presentation is easily converted into percentiles. Because of the similarities, there is no reason to present the individual analyses; however, the regression coefficients for the subsets of the data set are found in Appendix B of this article.

Glass and Smith closed with this statement: "Taking all findings of the meta-analysis into account, it is safe to say that between class-sizes of 40 and one pupil lie more than 30 percentile ranks of achievement... There is little doubt that, other things equal, more is learned in smaller classes" (pp. 45-46).

Commentary Regarding the Glass and Smith Study

Recall the reason for including the parabola (S^2) and the (L-S) term in the regression equation was presented by Glass and Smith (p. 17) as follows:

> The regression model selected accounted for variations in Δ (S-L) by means of S, S^2 and L. Obviously, something more than a simple linear function of S and L was needed, otherwise a unit increase in class-size would have a constant effect regardless of the starting class-size S; and the S^2 term seemed as capable of filling the need as any other. The size differential between the larger and smaller class, L-S was used in place of L for convenience.

The reason for the consistency property was presented as:

> The problem now is to find the set of $\hat{\Delta}$'s in this surface that can be depicted as a single curved-line relationship in a plane. The property that must hold for a set of Δ's before they can be depicted as a connected graph in a plane is what might be called the consistency property [underline in original]:
>
> $$\Delta_{n1\text{-}n2} + \Delta_{n2\text{-}n3} = \Delta_{n1\text{-}n3}\ldots$$

This section reviews whether the terms S^2 and (L-S) were appropriate choices; whether there are unintended consequences of these choices; and whether the inclusion of the consistency property transformation was warranted.

The Glass and Smith regression equation can be graphed as a two dimensional curve when L is set to a fixed value.[8] The U-shaped curve (parabola) is depicted in Figure 2 with and without the (L-S) term, and the (L-S) term is depicted separately. The value for the large class size is set at 65.81, so (L-S) will equal 0 at the right-hand portion of the graph.

From the presentation of the Glass and Smith results and the graph above, six questions or inconsistencies emerge:

(1) What is the interpretation of the relationship between achievement and class-size? The interpretation of the class size from the graph above seems obvious: As class size changes so does the level of predicted achievement, measured in Z-scores. Of note, the class sizes of 1 and 65.81 predict the same achievement level, with the lowest achievement predicted for a class size of about 33. This is because the S^2 term in the regression equation forms a U-shaped parabola. This representation does not correspond to the conclusion reached by Glass and Smith who report the regression results only to a class-size of 30.

(2) What was the reason for introducing the parabolic curve into the regression equation? Glass and Smith assumed that the relationship between achievement and class-size was nonlinear, and "... the S^2 term seemed as capable of filling the need as any other" (p. 17). No other rationale was provided. The reanalysis section of this paper will explore other options.

(3) What is the interpretation of the relationship between achievement and the (L-S) term? The achievement variable is related to the interval between the large and small class size (L-S). For example, if L = 65.81 and S = 1, then (L-S) = 64.81, with the coefficient of .00082, achievement is predicted to be an additional .053. The (L-S) term adds the most achievement when the class-size is 1 and gradually reduces as class size moves to 65.81, where no achievement is added. In other words, for every pupil added to the classroom, achievement decreases by .00082.[9] In order to make the two dimensional calculations, L must be a fixed value.

(4) What happens if the large class size (L) is set to another value? The value of L determines the relationship of the (L-S) line to the Z-axis (Z-score of 0). If L is set to a lower class size, the (L-S) line shifts lower and, as a consequence, the parabolic curve also shifts lower. In Figure 1, the (L-S) line intersects the Z-axis at a class size of 65.81 because of the value set for L was set at 65.81. Setting a different value to L does not change the basic relationship, only the magnitude of the Z-score; and because the coefficient is small, the magnitude of change is small. The value of L would be important, however, if the regression equation was linear (no S^2 term). In that case, L should be set to the average class-size where the achievement value would also be at the average—a Z-score of 0.

(5) What is the consistency property, and is it necessary? The consistency property transformation is offered for two reasons. Reason one is that the whole must equal the sum of the parts, or the sum of the intervals A to B and B to C must equal the interval A to C. Glass and Smith provided no illustration or example of why the condition was not met in the regression equation and, therefore, the necessity for a transformation. The conditions of the consistency property are met in the presentation of the regression results. (See Table 1.) Moreover, there is no necessity to apply the consistency property transformation to any linear or parabolic relationship. The line and the parabola are in a mathematical class called polynomials, which are continuous functions within the closed interval of the data points; the consistency property is inherent. In all circumstances, the Z-value for the intervals S_1 to S_2 plus the Z-value for the interval S_2 to S_3 equals the Z-value for the interval S_1 to S_3. Therefore, no transformation was necessary. (See also, Appendix A.)

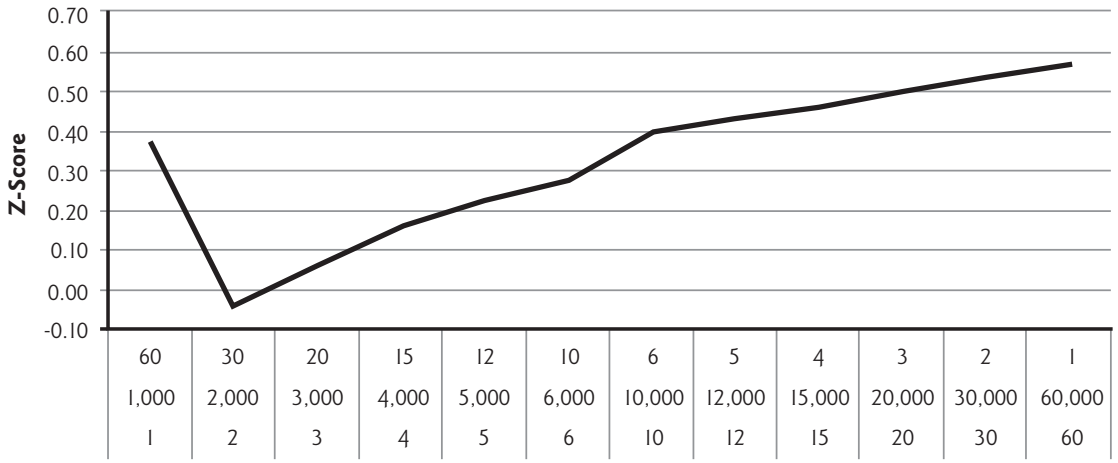
Glass and Smith (p. 18) proposed that the second reason for the consistency property transformation was to produce a "single line curve in a plane" from a three-dimensional surface. Apparently, they assumed the consistency property was related to the (L-S) term and considered it a third dimension. The transformation via the consistency property was not necessary to change a three-dimensional surface into a two-dimensional plane. The change is accomplished by setting L to a fixed value; indeed, setting a value for L is the only way to establish the connected curve in a plane.

The transformation via the consistency property made a fundamental change in the relationship between predicted achievement and class size. Up to P= 23, S was a variable; L was fixed; and the transformation was not applied. Above P = 23, the transformation was applied; S was fixed; and L was the variable. In essence, the transformation was only for values above S = 23 (P= 23). If the whole equals the sum of the parts below 23, then the whole equals the sum of the parts above S = 23, and the transformation is not necessary. If the relationship between achievement and class size is two dimensional below a class size of 23 (by setting the value of L), then it is two-dimensional above 23 (by setting the value of L). The value of L is immaterial to the number of dimensions. Glass and Smith's reasoning is not compelling; their logic is mathematically suspect, i.e., interchanging the character of S and L between fixed and variable.

The parabolic curve was an acceptable solution for class sizes between 1 and 23 because it was consistent with generally held perceptions. Because the parabolic curve was not consistent with perceptions for class-sizes above 33 (the low point), a method was employed that maintained the perceptions and modified the equation, hence the consistency property transformation. It appears that the consistency property transformation was invoked to reconcile the fact that the regression curve moves upward from the minimum and continues upward for all values of small class size, which extend well beyond 66. The value 65.81 is the class size where achievement is virtually the same as a class-size of 1. Essentially, it appears that the consistency property transformation was invoked to avoid this dilemma. If the S^2, the (L-S) terms, and the consistency property were not included in the methodology, there would be no dilemma.[10]

(6) Why are nearly all the value of the Z-scores above zero, when one would expect about half the values to be below the standard score mean of zero? The predicted Z-score values are mostly always above zeros because of the parabola and the consistency property;

Figure 3
Cost Implications of Reducing Class Size: Glass and Smith Regression Equation



in other words, decisions by Glass and Smith. Under normal circumstances, one-half of the observations will be negative; or half will be below average. The reanalysis section of this article will address this issue more specifically.

Class Size or Staff Adequacy?

In the methods section, Glass and Smith discussed the difference between class size and staff adequacy, and provided their reasons for choosing the first for the analysis. No discussion was entertained regarding the potential value of teacher aides, specialized teachers, or administrators as alternatives to increasing the number of classroom teachers. Perhaps there is a way to determine a cost-effective mix of these various educational roles (Phelps 2008).

The staff adequacy measure highlights the number of teachers required to achieve a particular class size, thus shedding light on the potential cost of reducing class size in relationship to increased achievement. For 66 students, it would take only one additional teacher to reduce class size from 66 to 33, for a total of 2 teachers, but it would take an additional 64 teachers to bring class-size to 1. Clearly class size and staff adequacy are on different measurement scales. It is possible to convert the class size ratio, the number of students (S) in a class with one teacher (T), or 1/S, to a measure of staff adequacy (the number of teachers (T) for a given number of students (NS), or T/NS), or 1/S = T/NS). For example, if the class size is 4 (1/4), and the number of students was set at 60 (NS), then 1/4 = T/60 = 15 /60; that is, it would require 15 teachers to have a class size of 4 for 60 students.

When the Glass and Smith regression curve is converted to the staff adequacy measure based on 60 pupils, the cost implications become clear. As class size is reduced, there is an increased cost per pupil (based on $60,000 per teacher) because of the increased number of teachers. As class size is reduced, the predicted achievement does increase (above a class-size of 30), but only up to a point, at which it levels off. Notice the different increments of teachers presented in Figure 3. Initially, class size is reduced dramatically with the addition of 1 teacher. After 10, the number of teachers must increase substantially to reduce class size; the last increment requires 30 additional teachers.

Observations Regarding Glass and Smith

Several initial questions were raised upon looking at the Glass and Smith regression curve. To follow are four observations based on the commentary above.

(1) Why are the relationships all above the 50th percentile? Glass and Smith made a reasonable decision to establish the 50th percentile as the reference point absent any other persuasive point. However, for any distribution only half of the observations can be above the 50th percentile. Their decision creates a strange world where every class size predicts above average achievement. It is logically inconsistent. Is there another way to interpret the situation? The reanalysis in the fifth section of this article addresses this issue.

(2) Why is the relationship curved? Glass and Smith included a squared term in the regression equation because they assumed the relationship between achievement and class size was curved, and the parabolic curve "seemed as capable of filling the need as any other" (p. 17). What is illustrated in the Glass and Smith figure is essentially the left side of the parabolic curve. The right-hand side was modified via the consistency property transformation. Is it possible that the relationship between achievement and class size is not parabolic? The purpose of the reanalysis will be to determine the natural shape of the curve.

(3) Why are the relationships for class sizes above 40 not reported? Glass and Smith used a consistency property to reformulate the original regression equation. The effect of the reformulation was to change the right side of the parabolic curve to avoid the dilemma of having large class sizes predict achievement at the same level as small class sizes. The purpose of the reanalysis will be to account for the full range of data and to address this dilemma.

(4) How many teachers are necessary to reduce the class size from 60 to 1? Figure 2 provides a general idea. Importantly, the measurement scale used in representing the Glass and Smith findings is not an equal interval scale with respect to the number of teacher required to achieve the respective class sizes. The number of teachers and the associated cost of reducing class size increase geometrically. For what class size range might it be cost effective to make the investment? The reanalysis will consider this issue.

Reanalysis of Glass and Smith

When commenting on the Glass and Smith study, two of their methods were questioned: (1) the inclusion of the S^2 and (L-S) terms in the regression equation; and (2) the application of the consistency property transformation.

When discussing the possible analytical methods, given the data available from different studies, Glass and Smith (p. 17) stated: "A few moments reflection will reveal that there is no obvious or simple way to connect these values into a single connected curve." This section tests this statement by proposing another way to connect the data values into a single connected curve. If the results from this other way and Glass and Smith methodology are essentially the same, then their findings will be confirmed. If, however, the results are not the same, then the reader will have to judge the validity of the two approaches and the plausibility of the different results. The purpose of this reanalysis is to identify the relationship between achievement and class size without relying on the questioned methods.

Mathematical Analysis

Glass and Smith provided three critical pieces of data for the reanalysis: (1) the difference in achievement between the smaller and larger classes, measured in Z-scores ($\Delta Z_{(S-L)}$); (2) the small class size (S); and (3) the large class size (L). If the smaller class size has the larger Z-score, the value of the outcome measure is positive, and vice versa. However, $\Delta Z_{(S-L)}$ is not the desired achievement variable for the analysis: Z_S is the desired variable. From these data, the object of this reanalysis is to find a function other than the one presented by Glass and Smith predicting the value of Z for the entire data range of class sizes:

$$Zcs = f\ (CS)$$

The strategy of this reanalysis is to convert each observation from the data set into points on a line segment defined by Z and each class size between S and L. Where the class size points on the line segments are in common, the Z's are averaged. The averages for each class size point are then joined over the full range of class-sizes forming a data-driven curve.[11]

Figure 4
The Relationship Between Achievement and Class Size Based on Reanalysis: Data-Driven Curve

In the section, "Describing the Class-size and Achievement Relationship," Glass and Smith concluded (p. 17), "...various values of $\Delta_{(S-L)}$ arising from different studies can show confusing inconsistencies." This is because various $\Delta_{(S-L)}$ span different ranges of class size. When $\Delta_{(S-L)}$ is divided by (L-S), the inconsistencies disappear. With this value ($\Delta_{(S-L)}$ / (L-S)), a separate value can be calculated for each class size within the range. For example, instead of a single observation for $\Delta_{(10-20)}$, there can be 11 observations—one each for class size, starting with 10 and continuing through 20. With this shift in the paradigm, changing the achievement variable to a Z-score, the necessity for (L-S), S^2, and the consistency property all disappear. This paradigm seems obvious and is clearly less complex.

We start with the definition of the measure of achievement outcome:

$$\Delta Z_{(S-L)} = Z_S - Z_L$$

The achievement measure $\Delta Z_{(S-L)}$ is divided by the difference in the class sizes, $CS_L - CS_S$, to obtain the slope (M):

$$Z_S - Z_L / CS_L - CS_S = M$$

Therefore, the line segment between SS - SL is:

$$Z_{CS} = M\ CS_{(S-L)} + B$$

where B is the Z-axis intercept. The interpretation of this function is straightforward: For any give value of class size (CS), there is a corresponding value of Z_{CS}, measured as an achievement Z-score. If achievement levels decrease as class size increases, the slope is negative. Conversely, if achievement levels increase as class size increases, the slope is positive. Therefore, the sign of the achievement variable in this context is the opposite of the sign of the achievement variable in Glass and Smith.

With this slope-intercept line function, a new analytical paradigm emerges. The slope for each observation is calculated and a Z-score recorded for each class-size within the line segment. These Z-scores are averaged rather than summarized by a least-squared method because there is no intent to make statistical inferences. By joining these Z-scores into a line, a representation of the relationship between achievement and class-size is obtained directly, independent of any predetermined decisions of the researcher. In contrast, Glass and Smith relied upon the predetermined parabolic function, the (L-S) term, and a consistency property.

The relationship between achievement and class size with the method proposed in this reanalysis can take on any shape—linear, curved, or a combination—and accommodates positive and negative slopes. Using the above interpretation, 40% of the observations in Glass and Smith's data set had positive slopes. If these observations were clustered together in one region of class sizes, there would be a corresponding upswing in the curve. This method also allows for an inspection of the relationship between achievement and class size to determine if it is nonlinear in some ranges and what might be the appropriate curve to fit via future regression analysis.

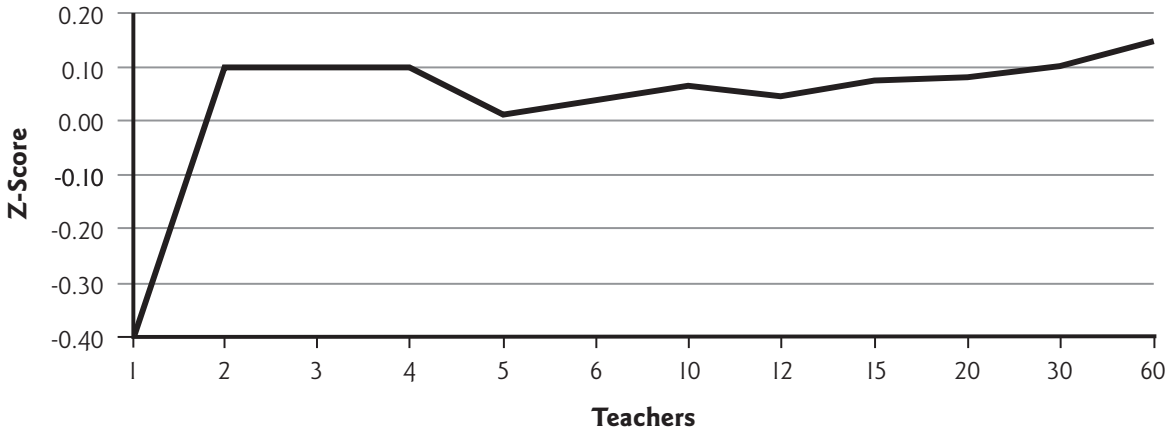
Data Set for Reanalysis

Although Glass and Smith's raw data were listed in an appendix to their study, it is not available in a current electronic format.[12] As a result, the data for this reanalysis were entered by hand from the appendix, but not all data were included. Only data for the categories of elementary school classes (all subjects combined), reading, mathematics, and language were transcribed while the data for the categories of psychology, natural/physical sciences, social sciences and history, and "all others" were excluded. This decision was made for two reasons: First, transcribing was labor intensive; and, second, the categories of elementary school classes, reading, mathematics, and language were considered to be the more relevant subjects in reviewing public school achievement.

Glass and Smith included 725 comparisons taken from 77 studies, including 343 observations for elementary school, 39 in reading, 84 in mathematics, and 144 in language. For the reanalysis data set, there were 309 observations for elementary school, 21 in reading, 84 in mathematics, and 50 in language.

While entering the data, some discrepancies were observed. There were data for the number of pupils and the number or teachers for most of the observations as well as an entry for the ratio of the number of pupils per teacher, but they did not always align. For example, the first data entry for the smaller class size showed 60 students for 10 teachers but with a ratio of 1 instead of 6. There was no way to know the reason for the inconsistency, but because the actual numbers were available, it seemed logical to enter the newly calculated figure rather than the suspicious ratio. This principle was applied to other similar observations. In addition, there was a series of entries with the number of pupils but no entry for the number of teachers. At the same time, the ratio was always

Figure 5
Cost Implications of Reducing Class Size: Data-Driven Curve



1. These observations were not included in the reanalysis because there were a substantial number of observations with a small class size of 1 that could be used.

From the reanalysis data set, the slope for each of the observations was calculated and inspected. Four observations had slopes substantially higher or lower when compared to the rest of the data set. These four inconsistent observations were considered extreme outliers and eliminated from the reanalysis. As a result of these decisions, a total 463 observations comprised the reanalysis data set.

What was left was to decide was the value of B, the Z-score intercept. Because the achievement variable was measured in Z-scores, with the midpoint or average at zero, B could be set so that the average class size would correspond to a Z-score of zero. This method of estimating B is not perfect, but it gives some indication of the relative contribution class size makes to achievement over the full range of class sizes. It also avoids the dilemma of having all class sizes predicting above average achievement. The result of the reanalysis is portrayed in Figure 4.

The representation of the data-driven curve presents a more complicated picture of the relationship between achievement and class size than that of the Glass and Smith regression curve. The data-driven curve is essentially U-shaped between 1 and 33, then consistently downward to 75. The predicted achievement level at a class size of about 33 is higher—almost double—than the achievement level at a class size of 1. However, the similarity of predicted achievement between class sizes of 1 and 65.81 is not present, as was the case with Glass and Smith. The substantial number of positive slope observations concentrated between class sizes 15 and 33 explains the upward curve.

From a class size of about 33 upward, there was a continuous and consistent reduction in predicted achievement. The anomalies in the curve at a class size of 54 and between 56 and 60 were due to slopes that are substantially different from the corresponding studies.[13] Removing these observations from the data set would smooth out the descending line.

Figure 6
Comparison of Four Relationships Between Achievement and Class Size

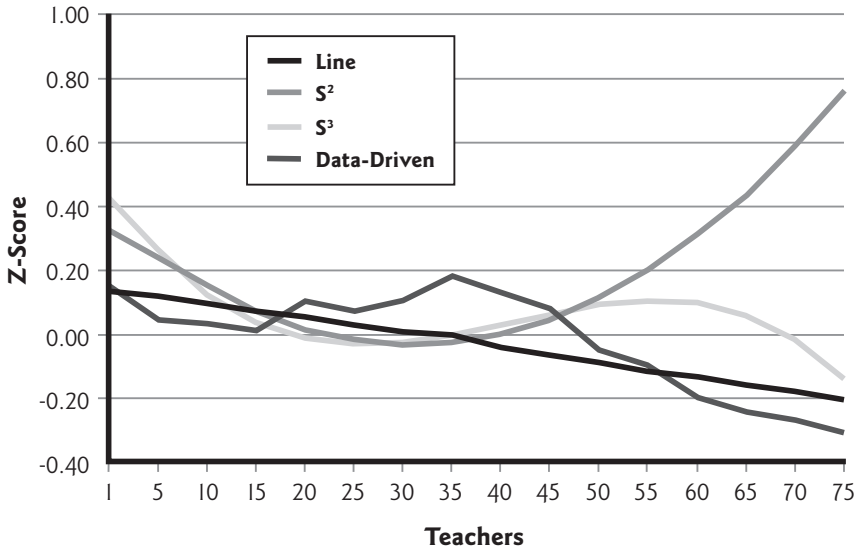


Figure 5 depicts the number of teachers required for 60 pupils in relation to the predicted achievement level. As teachers are added, so does the predicted achievement, moving from one teacher to two, or class size from 60 to 30. However, there is a point where the increase in predicted achievement does not warrant the increase in the number of teachers and associated cost. The policy implications derived from the reanalysis portrayed in this graph are different than those from the staff adequacy transformation of Glass and Smith found in Figure 3.

The cost implications from Figure 5 are straightforward. Moving from a class size of 60 to 30 would require an additional teacher, from one to two, essentially doubling the cost. However, there would be a substantial gain in predicted achievement largely justifying the increased cost. But moving from a class size of 30 to 1 would require another 58 teachers with the amount of achievement gain largely uncertain.

Conclusions

The generalizations made in this section were based on a subset data from the Glass and Smith study. The conclusions were reached by comparing the curves generated using the Glass and Smith regression methodology with the data-driven curve methodology used in the reanalysis. No attempt has been made to include data, findings, or conclusions from other class size research.

In the graph below, four relationships between achievement and class-size are depicted, all based on the revised data set. (See Figure 6.) Three are based on Glass and Smith's regression analysis, and the fourth is based on the reanalysis. The first relationship removes the S^2 term from the Glass and Smith regression equation to form a line; the second, the original equation, includes an S^2 term producing a single-bend curve (parabola); the third includes a S^3 term adding another critical point producing a double-bend curve; and the fourth is the data-driven curve. The three regression curves are continuous curves, so the consistency property transformation is not applied for the reason provided earlier. (See Appendix B for regression coefficients and statistics.)

As can be seen in Figure 6, the line is the most straightforward representation of the relationship between achievement and class size. Predicted achievement decreases as class size increases. The line is inconsistent with the data-driven curve, especially for class sizes in the range of 15 to 35. The single-bend curve (the Glass and Smith regression curve) predicts achievement to decrease as class size increases to about 33, at which point the interpretation becomes counterintuitive—achievement increases.[14] This curve does not resemble the data-driven curve or the linear representation. The double-bend curve suggests a complex relationship between achievement and class size. It somewhat resembles the data-driven curve, but in a different phase. In each of the cases, a problem of interpretation arises:

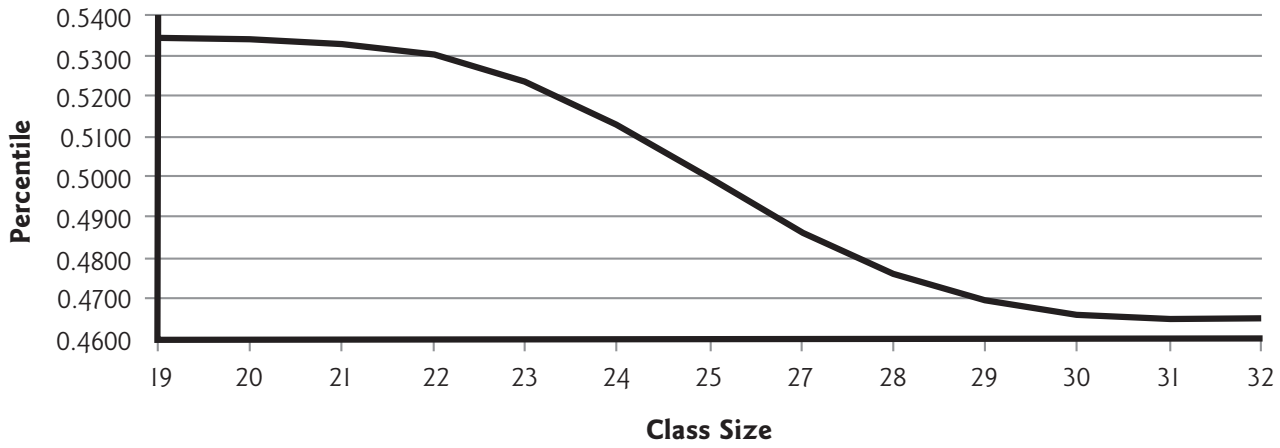
- The line and all curves indicate a gain in estimated achievement as class size moves smaller than about 15.
- There is a predicted gain in estimated achievement as class size moves larger than about 15 for the data-driven curve and about 30 for the two regression curves. The line does not indicate a gain.
- The data-driven curve indicates a drop of estimated achievement as class size moves larger than about 35 while the double-bend curve indicates a drop in achievement as class size moves larger than about 55.

What conclusions can be reached given these indications? The single-bend curve is not supported by the evidence of the data-driven curve or the double-bend curve. While the evidence tends to support the notion that achievement would increase for class sizes smaller than 15, the evidence also supports the notion that such class size reductions are cost prohibitive. The evidence supports the notion that class sizes over a certain size are associated with a decrease in achievement; the exact critical point is in doubt based on these data and analyses. In contrast, the evidence does support lowering class sizes from the large extremes, and there are indications that the potential gain would offset the marginal cost. The influence of class size between about 15 and about 45 is unclear, other than the general conclusion that the relationship between achievement and class size is indeed complicated as Porwell (1978) suggested.[15]

Representing the Relationship between Achievement and Class Size Based on a Normal Curve

If one would make some basic assumptions regarding a class size curve, what would those assumptions be? First at the larger

Figure 7
Theoretical Relationship between Achievement and Class Size: Hypothetical Curve



class sizes, it would be fair to assume that by adding one student to a class of 100 students, there would be little if any difference in achievement. With this assumption, a well-matched curve would show gradually decreasing achievement as class size increased approaching a lower bound; i.e., a lower-bound asymptote. Second, at the smaller end of class size, it would be fair to assume that the difference in achievement by removing one student from a class of 5 students would show gradually more achievement as class size decreased approaching an upper bound of 1; i.e., an upper-bound asymptote. Third, it would be fair to assume that the average class size would predict the average achievement level. Finally, it would be fair to assume that all class sizes above the average would predict achievement below the average and vice versa for class sizes below the average predicting achievement above average. These assumptions address the difficulties with the data-driven and two-bend curves presented previously.

There is a curve meeting these conditions. This curve has its roots in normal curve statistics and provides a more reasonable explanation than the other curves. The details are explained fully elsewhere (Phelps, 2008). In summary, the amount of variance explained by a regression equation can be converted to the curve in Figure 7. With the dependent and independent variables measured in standard scores (Z-scores), the amount of variance explained (R^2) can be converted into a normal curve with the same area. When the normal curve is integrated (cumulative area under the curve), the result is an S-shaped curve, asymptotic at the upper and lower bounds, with the average class size predicting the average achievement.

Determining the amount of variance explained by class size is complex because class size is likely to be correlated with other important variables such as socioeconomic status (SES), expenditures, teacher qualifications, support staff, and instructional materials. Studies with these variables could provide estimates of possible ranges of the variance attributable to class size; these estimates can be instructive in policy decision-making (Phelps, July 2008). While the data set from the Glass and Smith study is not suitable for this type of analysis, at least an example can be offered. This example has an average class-size of 25, a standard deviation of 2, and an R^2 of .07 (the average R^2 of the three regression curves is .07).

In reality, class size does not range from 1 to 70, as does the data set, but is more likely to be in the range suggested above. More likely, the curve has a consistently downward slope. It would seem that the likely relationship between achievement and class size is more similar to the curve suggested in Figure 7 than the complex curves depicted in Figure 6.

In summary, there is a likely relationship between class size and achievement, but the relationship is exceedingly complex. At the same time, the financial cost of reducing class size as a primary method of increasing achievement is not warranted. The conclusion to be drawn from these three points is that the substantial influence of Glass and Smith (1978) in changing policy related to class size was/is probably unwarranted. In the final analysis, the class size policy question comes down to what is believed and what is accepted. Does one believe in the analytical results and accept the methodology, or does one believe in the methodology and then accept the results?

References

Glass, Gene V., and Mary Lee Smith. *Meta-Analysis of Research on the Relationship of Class-size and Achievement*. San Francisco, CA: Far West Laboratory for Educational Research and Development, 1978.

Phelps, James L. "Optimizing Educational Resources: A Paradigm for the Pursuit of Educational Productivity." *Educational Considerations* 35 (Spring 2008): 3-18.

_____. "Measuring and Reporting School and District Effectiveness." *Educational Considerations* 36 (Spring 2009): 40-52.

Porwell, P.J. *Class Size: A Summary of Research*. Arlington, VA: Educational Research Service, Inc., 1978.

Endnotes

[1] All subsequent references to Glass and Smith in this article refer to Gene V. Glass and Mary Lee Smith, *Meta-Analysis of Research on the Relationship of Class-Size and Achievement* (San Francisco, CA: Far West Laboratory for Educational Research and Development, 1978).

[2] Meta-analysis is a research method that takes data from many individual studies and combines them into a new analysis.

[3] This is a curious statement. What if the situation were shifted to a supermarket where the price of potatoes was 1 lb. for 70 cents, 3 lbs. for $2.00, 5 lbs. for $3.10, 10 lbs. for $6.00, and 25 lbs. for $13.00? What would the shopper do?

[4] In the previous section, emphasis was added to three points. These points are critical in later portions of this paper: (1) No obvious and simple alternative; (2) including the S^2 term; and (3) including (L-S) term.

[5] See Appendix A of this article for a detailed discussion of the consistency property.

[6] The sum of the intervals should equal 0, as it does considering the rounding error.

[7] The differences in the values in the Table 1 are due to a different value being set for the large class size.

[8] The equation can be graphed in three dimensions with L being the third, starting with 1 and continuing to the largest class-size in the data set. To determine a point on the surface, an arbitrary value for L must be selected in order to evaluate (L-S).

[9] While the relationship between achievement and class-size—the S variable—is parabolic, the relationship between achievement and (L-S) is linear.

[10] When Glass and Smith added a squared term to their equation representing the relationship between class size and achievement, they applied the same mathematical function used to describe a thrown ball—a parabola. So whether intended or not, their class size curve and a thrown ball should follow the same general path. If their parabola assumption were based on fact rather than supposition, and if their consistency property were mathematically correct, then by mathematical symmetry, a thrown ball would follow the upside-down Glass curve (Figure 1) and would never

come down! Conversely, if the thrown ball path is correct, then their squared term assumption, their consistency property, or both, are faulty.

[11] For example, to find the best price per pound, divide the price by the number of pounds. The shopper determined the cost per pound in cents was 70, 67, 62, 60, and 52. These numbers can be placed into a curve depicting the price per pound for various packaging weights.

[12] Author's correspondence with Gene Glass.

[13] See observation #369, study #55, and observation #373, study #4.

[14] Achievement at class-size 1 and 61 (rather than 65.81) is the same because of the change in the data set.

[15] The data-driven curve generated by the reanalysis is complicated to explain; that is, why is a class-size of 33 be the best level for achievement? One must take into consideration that the data in the reanalysis may not be representative, and hence other data sets and other paradigms should be used to test the underlying question.

APPENDIX A

Discussion Regarding the Consistency Property

Glass and Smith (1978, 17) stated [italics added for emphasis]:

Fitting this model by least-squares will result in the curved regression surface:

$\hat{\Delta}_{(S-L)} = \hat{\beta}_0 + \hat{\beta}_1 S + \hat{\beta}_2 S^2 + \hat{\beta}_3 (L-S)$

The problem now is to find the set of $\hat{\Delta}$'s in this surface that can be depicted as a single curved line relationship in a plane. The property that must hold for a set of $\hat{\Delta}$'s before they can be depicted as a connected graph in a plane is what might be called the consistency property:

$\Delta_{n1-n2} + \Delta_{n2-n3} = \Delta_{n1-n3}$

for n1 < n2 < n3. *If this property is not satisfied, then one is in the strange situation of claiming that the differential achievement between class-size 10 and 20 is not the sum of the differential achievement from 10 to 15 and then from 15 to 20.*

When the consistency property is imposed on [regression equation] *(2),* it follows that:

$\hat{\beta}_0 + \hat{\beta}_1 n_1 + \hat{\beta}_2 n_1^2 + \hat{\beta}_3 (n_2-n_1) + \hat{\beta}_0 + \hat{\beta}_1 n2 + \hat{\beta}_2 n2^2 + \hat{\beta}_3 (n_3-n_2) =$
$\hat{\beta}_0 + \hat{\beta}_1 n_1 + \hat{\beta}_2 n_1^2 + \hat{\beta}_3 (n_3-n_1)$ (3)

Simple algebraic reduction produces the following:

$\hat{\beta}_0 + \hat{\beta}_1 n_2 + \hat{\beta}_2 n_2^2 = 0$

The two solutions to the quadratic equation...are points n_2 such that the $\hat{\Delta}$ is measured with n_2 as either the larger, L, or smaller, S, class size then the resulting set of $\hat{\Delta}$'s will lie on the four dimensional regression curve...but can be depicted as a single line curve in a plane. Since n_2 becomes the point around which values of n_1 and n_3 are selected, it will be called the pivot point [emphasis in original]. That there are two solutions for n_2 is perplexing; fortunately in the analyses to be reported the two corresponding curves were virtually parallel in practice.

A single line curve in a plane can be constructed by solving for one or the other values of n_2 in (4) and constructing a set of $\hat{\Delta}$'s values. *These values will give the standardized mean differences in achievement between n_2 and any other class size.* The curve that connects these $\hat{\Delta}$'s has no non-arbitrary starting point. *One can assume for convenience sake that the achievement curve (z), instead of the differential achievement curve ($\hat{\Delta}$)* is centered around an arbitrary class size, e.g., something like the national average in the low 20s (pp. 17-19).

The purpose of this discussion is to test the assumptions underlying the consistency property as described above. (Note the italicized passages.)

1. Under what circumstances is the differential achievement between class size 10 and 20 the sum of the differential achievement from 10 to 15 and then from 15 to 20?
2. Can the consistency property be logically imposed on the regression equation?
3. If the consistency property cannot be logically imposed on the regression equation, is there an alternative formulation?
4. What is the nature of the achievement variable? The achievement variable in the data set is $\Delta_{(S-L)}$, but why has the interpretation changed to a Z-score after the regression coefficients have been applied to the equation?
5. What are the consequences of the alternative formulation?

In order to critique the "imposition" of the consistency property (equation (3)) on the regression equation (equation (2)), three achievement values must be obtained—one each for three sequential and equidistant class sizes (e.g., class-sizes of 10, 15, and 20 as suggested). For the critique, the selected coefficients values are: $\beta_0 = 2$, $\beta_1 = -.1$, $\beta_2 = 0$, and $\beta_3 = .01$. These values have been set to make the calculations simpler and clearer (eliminating the squared term making the relationship linear). The selection of the values does not affect the underlying principles or conclusions. Substituting these values, regression equation (2) becomes: $\Delta = 2 - .1S + .01(L-S)$. The consistency property in equation (3) can be expressed as three equations where the sum of the first two equals the third ($\Delta 1 + \Delta 2 = \Delta 3$):

$\Delta 1 = \beta_0 + \beta_1 n_1 + \beta_2 n_1^2 + \beta_3 (n_2-n_1)$ or $2 - .1 * 10 + .01(15-10) = 2 - 1 + .05 = 1.05$
$\Delta 2 = \beta_0 + \beta_1 n_1 + \beta_2 n_1^2 + \beta_3 (n_2-n_1)$ or $2 - .1 * 15 + .01(20-15) = 2 - 1.5 + .05 = 0.55$
$\Delta 3 = \beta_0 + \beta_1 n_1 + \beta_2 n_1^2 + \beta_3 (n_2-n_1)$ or $2 - .1 * 10 + .01(20-10) = 2 - 1 + .1 = 1.10$

The algebraic reduction of the equations (3) becomes:

$\beta_0 + \beta_1 n_2 = 0$ or $\beta_0 = - \beta_1 n_2$ (4)

Equation (3) is false ($\Delta 1 + \Delta 2 \neq \Delta 3$). Also, equation (4) is false ($2 + (-.1 * 15) \neq 0$). Equation (3) will be true only when $n2 = - \beta 0 / \beta_1$, or -2/- .1, or a class size of 20 which contradicts the initial condition of n2 = 15. The equations proposed by Glass and Smith for meeting the consistency property conditions are unsatisfactory. The task is to identify a workable alternative formulation.

The solution to consistency property equations will be clearer if the regression equations are graphed. Graphing the expression $\Delta = \beta_0 + \beta_1 S$ is straightforward: the expression is represented by a line with a slope of -.1 and the Δ intercept of 2 (at S = 0, Δ = 2). Graphing the expression $\Delta = \beta_3 (L-S)$ is problematic; while the slope is .01, there is not a consistent intercept. For Δ, the intercept is 15 (when S = 15,

(L-S) = 0 and Δ = 0). For the other two equations the intercept is 20. In other words, L is the intercept, and it is not the same in each equation. As a result, the (L-S) term produces a family of lines and not a single line, as with the other expression. This difference between the two expressions is critical.

Looking for an alternative, there are two primary criteria: (1) Δ1 + Δ2 must = Δ3; and (2) because the equations are linear (by setting the squared term to 0) and the class-sizes are sequential and equidistant, the values of Δ1, Δ2, and Δ3 must also be sequential and equidistant.

In the first test for an alternative, the large class size is set to a fixed value (L=20), and Δ3 is calculated with the value of the third class size:

$\Delta 1 = 2 - .1*n_1 + .01*(L-n_1)$ or $\Delta 1 = 2 - .1*10 + .01*(20-10) = 1.10$

$\Delta 2 = 2 - .1*n_2 + .01*(L-n_2)$ or $\Delta 2 = 2 - .1*15 + .01*(20-15) = 0.55$

$\Delta 3 = 2 - .1*n_3 + .01*(L-n_3)$ or $\Delta 3 = 2 - .1*20 + .01*(20-20) = 0.00$

Again, Δ1 + Δ2 ≠ Δ3! However, Δ1, Δ2, and Δ3 are sequential and equidistant. The situation does not change if L is set to another value, although Δ1 + Δ2 does = Δ3 at L = 291. But if any of the class-sizes change, so does the value of L; so there are an infinite number of solutions to the equations! Interestingly, the average class size must be 20, for when S =20, achievement is predicted to be 0, and the average class size equals the average achievement (a Z-score of 0). In order to evaluate the regression equation, L must be set to a constant to preserve a consistent relationship among the class-sizes. The achievement variable is not measured in terms of Δ and/or the formulation is incorrect in that the whole is not the sum of the parts but is correct in that the values are sequential and equidistant. Equation (2) is true. Even with the change, equation (3) is not true.

For the second test for an alternative, the achievement variable is assumed to be Z-scores, and the Δ is assumed to be the difference between two Z-scores, or: Δ1 = (Z2 – Z1), Δ2 = (Z3 –Z2), and Δ3 = (Z3-Z1), or $(f(S2) - f(S1)) + (f(S3) - f(S2)) = (f(S3) - f(S1))$.

$Z1 = 2 - .1*n_1 + .01*(L-n_1)$ or $Z1 = 2 - .1*10 + .01*(20-10) = 1.10$

$Z2 = 2 - .1*n_2 + .01*(L-n_2)$ or $Z2 = 2 - .1*15 + .01*(20-15) = 0.55$

$Z3 = 2 - .1*n_3 + .01*(L-n_3)$ or $Z3 = 2 - .1*20 + .01*(20-20) = 0.00$

Substituting, (.55- 1.10) + (.00 - .55) = (.00 – 1.10) or (-.55 - -.55) = -1.1. Both criteria are met. Therefore, when L is set to a fixed value, the achievement variable is measured in Z-scores, and Δ is the difference between two Z-scores, "...the differential achievement between class-size 10 and 20 is...the sum of the differential achievement from 10 to 15 and then from 15 to 20" (p. 18). With this interpretation, the logical condition is met, and the regression equation is graphically portrayed not as a surface but as two lines which, when added together form a "single curved-line in a plane." This interpretation is consistent with the results presented in Table 1 using the actual regression equation.

Under Glass and Smith's overly-complicated consistency property formulation, the logical condition is not met. In practice, they do set L to a fixed value, but make other changes, which are discussed in this article. Based on this analysis, it is inappropriate to apply the consistency property transformation.

APPENDIX B

Table B-1
Regression Coefficients from Glass and Smith Meta-Analysis

Studies	*Intercept*	*S*	S^2	R^2
Elementary students	0.38503	-0.02995	0.00052	0.255025
Secondary students	0.75539	-0.05024	0.00071	0.192721
Poorly controlled	0.07399	-0.00587	0.00009	0.034969
Well controlled	0.69488	-0.06334	0.00128	0.385641
All	0.57072	-0.03860	0.00059	0.181476

Source: Glass and Smith (1978, 33, 39). R^2 calculated by author from multiple R.

APPENDIX C

Table C-1
Regression Coefficients, R^2, and Z-axis Intercepts from Reanalysis

	Intercept	*L-S*	*S*	*S^2*	*S^3*	*R^2*	*Z = 0*	*Z = 0*	*Z = 0*
Line	0.141156	0.002786	-0.004679			0.034	30.76		
S^2	0.356798	0.002891	-0.025273	0.000407		0.084	22.09	40.00	
S^3	0.461896	0.003502	-0.045211	0.001270	-0.000010	0.098	18.32	35.56	69.21

A Practical Method of Policy Analysis by Considering Productivity-Related Research

James L. Phelps

The basic notion underlying schooling is rather simple: Hire teachers to instruct students. From there, the tasks become more complicated. How many teachers should be employed? What assignments should the teachers be given, in the classroom or in a supporting role? What assistance should teachers receive from aides or volunteers? What role do administrators play? Schooling is even more than staffing: It includes the curriculum; methods of instruction, instructional materials, time of instruction, and home support including homework. All of these elements must combine into a unifying whole in order to achieve the desired educational goals. Goals other than achievement are important as well, e.g., staying in school, preparation for employment, and civic responsibility, just to name a few. However, the topic must be limited, so this discussion focuses only on the goal of student achievement.

Class size may be important in achievement, but it is not the only decision for policymakers. Class size plays a role, but the role is effectively fulfilled only when the other players are successful. Therefore, it is appropriate to address several questions: What goals are to be accomplished; what is the best distribution of personnel related to these goals; what roles do curriculum, instruction, time, and home support play; and how do the personnel work together effectively to achieve those goals? In the broadest sense, the fundamental question is: How are decisions made?

A Taxonomy of Class Size Decision Making

For the sake of discussion, three levels of decision making related to class size are presented. Generally speaking, there are three broad categories or levels:

(1) Professional and public opinion;
(2) A critical analysis of educational research evidence;
(3) A decision-making process, including: (a) establishing a set of clearly stated goals; (b) identifying a set of possible policy options to achieve the goals; (c) clearly stating the assumptions why each of the policy option would achieve the goals; and (d) evaluating each of the policy options to select the best alternative.

A case could be made that decision making based upon the first perspective is the most common. The premise of this article is to provide some rationale and ideas regarding how policymakers can move through the more sophisticated levels of the taxonomy—the critical analysis of educational research evidence and a structured decision making process. Undoubtedly, policymakers have intuitive answers to the complicated questions encompassing education, but the objective of good policymaking is to explicitly spell out those questions and underlying assumptions regarding the best answers.

- Will lower class sizes make a difference in student achievement? By how much?
- Will an increased number of other instructional staff have a beneficial impact on student achievement? By how much?
- Will effective instructional and organizational policies have a beneficial impact on achievement? If so, by how much?

The purpose of this discussion is to explore the policymaking process by exploring these issues through the research literature. The next article, "A Practical Method of Policy Analysis by Estimating Effect Size," further develops the issues raised here using data from Minnesota. The fourth article, "A Practical Method of Policy Analysis by Simulating Policy Options," suggests a method of policy analysis, based on the ideas and data from the previous articles, in order to investigate possible answers to the questions posed above.

This article is divided into three parts. In the first, Does Class Size Make a Difference: A Brief Overview of the Research,[1] a sampling of studies is presented. It should be noted that some research studies have included variables other than class size. The second section is titled, How Much of a Difference Does Class Size Make on Achievement? The 1978 meta-analysis of Glass and Smith suggested the possibility that achievement increases faster as class sizes become smaller. This study has influenced research and policy ever since. This section examines some of the issues concerning the nature of the relationship between class size and achievement. What is the magnitude of the relationship? What is the nature of the relationship, increasing as suggested by Glass and Smith, or some other pattern? This section notes that some other policy options might improve achievement either independently or in combination with lower class size. The third and final section closes with some observations.

Does Class Size Make a Difference? A Brief Overview of the Research Literature

Clearly, teachers and the public believe that small classes produce higher achievement. Whether their beliefs are supported by evidence is a separate question; nevertheless, beliefs have a major influence on the decision making process. Although the data are somewhat old, Robinson and Wittebols (1986) reported several polls indicating the magnitude of those beliefs. In most cases, lower

James L. Phelps holds a Ph.D. from the University of Michigan in Educational Administration. He served as Special Assistant to Governor William Milliken of Michigan and Deputy Superintendent in the Michigan Department of Education. Active in the American Education Finance Association, he served on the Board of Directors and as President. Since retirement, he spends a great deal of time devoted to music, composing and arranging, playing string bass in orchestras and chamber groups, as well as singing in two choirs. He resides with his wife, Julie, in East Lansing, Michigan.

class size was thought to be favorably associated with achievement, discipline, decreased drug use, decreased crime, and increased student motivation. There is little reason to think those beliefs have changed.

Hanushek (1989, 1998, 1999) has researched and written extensively on the issue of class size and its relationship to achievement.[2] He has offered evidence in four ways: (1) interpretation of historical aggregate data; (2) international comparisons; (3) econometric studies; and (4) analysis of controlled experiments. This overview follows the same structure.

Interpretation of Historical Aggregate Data

Substantially more teachers have been added to the U.S. system of education over time with little change in academic performance. Hanushek (1999) presented the changes in aggregate class size between 1960 and 1994, a reduction from about 27 to about 20. In contrast, the measure of achievement, NAEP (National Assessment of Education Progress), showed little change. The analysis went on to account for the changes in student population, changes in special education, and racial differences in achievement. Based upon his analysis, Hanushek (1999, 17-18) concluded:

> The available data and evidence suggest some uncertainty about the underlying forces related to families, school organization, class size, and achievement. Allowing for changes in family background and special education, however, it remains difficult to make a case for reduced class size from the aggregate data. A natural experiment in class size reduction has been going on for a long period of time, and overall achievement data do not suggest that it has been a productive policy to pursue. Nonetheless, the aggregate data are quite limited, restricted to a small number of performance observations over time and providing limited information about other fundamental changes that might affect school success (pp. 17-18).

International Comparisons

There is no systematic relationship between class size and achievement. The international analysis focuses on two examples. The first concerns the Third International Mathematics and Science study (TIMSS) for which the pupil-teacher ratios and achievement scores were correlated. The correlation was positive, higher ratios (more pupils in a classroom) were associated with higher performance, but thought to be a statistical artifact rather than persuasive evidence (Hanushek, 1998, 18).

The second analysis was a more systematic examination of international tests with 70 country-specific measures of pupil-teacher ratios and achievement. According to Hanushek and Kim (1995), the results were positive but statistically insignificant when controlled for parents' schooling. They added:

> Of course, there are many differences in schooling and societies of the sampled nations, so it would be inappropriate to make too much of these results. They do, however, underscore that the normal presumptions about the achievement effects of pupil-teacher ratio and class size are not found in the evidence (p. 19).
>
> Somewhat surprising, similar kinds of results are found if one looks across countries at the relationship between pupil-teacher ratios and student performance. While it is clearly difficult to develop standardized data across countries, to control for the many differences in populations and schools, and the like, there remains some appeal in looking across countries. The variation in class sizes and pupil-teacher ratios are larger than found within the U.S., leading to some hope that the effect of alternative intensities of teacher usage can be better understood. Even given the wide difference, there is no evidence that lower pupil-teacher ratios systematically lead to increased performance (p. 21).

In another study based on the TIMSS achievement measure, Woessmann and West (2002, 7) concluded:

> We estimate the effect of class size on student performance in 18 countries, combining school fixed effects and instrumental variables to identify random class-size variation between two adjacent grades within individual schools. Conventional estimates of class-size effects are shown to be severely biased by the non-random placement of students between and within schools. Smaller classes exhibit beneficial effects only in countries with relatively low teacher salaries. While we find sizable beneficial effects of smaller classes in Greece and Iceland, the possibility of even small effects is rejected in Japan and Singapore. In 11 countries, we rule out large class-size effects.

Econometric Studies

The number of econometric studies with statistically significant results are offset by an almost equal number of statistically insignificant studies. The econometrics studies are based on an input/output regression model controlled for socioeconomic status (SES)

Table 1
Distribution of Estimated Influence of Teacher-Pupil Ratio on Student Performance

School Level	Number of Estimates	Statistically Significant (%)		Statistically Insignificant (%)		
		Positive	Negative	Positive	Negative	Unknown
All schools	277	15	13	27	25	20
Elementary	136	13	20	25	20	23
Secondary	141	17	7	28	31	17

Source: Eric A. Hanushek, "The Evidence on Class Size," Occasional paper 98-1 (Rochester, NY: Wallis Institute of Political Economy, University of Rochester, 1998), 23, Table 5.

Table 2
Krueger's Re-Analysis of Hanushek's Meta-analysis

Results (in Percentages)	*Hanushek: Estimates Weighted Equally*	*Krueger: Estimates Weighted by Inverse of Number of Estimates in Study*	*Krueger: Estimates Weighted by Citation Frequency*	*Krueger: Estimates Derived from Regression Analyses of Original Estimates*
Positive and Statistically Significant	14.8	14.4	30.6	33.5
Negative and Statistically Significant	13.4	10.3	7.1	8.0
Statistically Insignificant	71.9	61.2	62.3	58.4

Source: Alan B. Krueger, "Understanding the Magnitude and Effect of Class Size on Student Achievement," in *The Class Size Debate*, edited by Lawrence Mishel and Richard Rothstein (Washington DC: Economic Policy Institute, 2002), 7, Table 1-2.

Table 3.1
Class Size and Student Achievement: Studies Clustered by Grade Level

Grade Level	*Total Number of Studies*	*Studies Favoring Small Class Size*	
		Number	*Percent (%)*
K-3	22	11	50.0
4-8	21	8	38.1
9-12	22	4	18.2

Source: Glen E. Robinson, and J.H. Wittebols, *Class Size Research: A Related Cluster Analysis of Decision Making* (Arlington, VA: Educational Research Services, Inc., 1986), 67.

Table 3.2
Class Size and Student Achievement: Studies Clustered by Reading Achievement

Grade Level	*Total Number of Studies*	*Studies Favoring Small Class Size*	
		Number	*Percent (%)*
K-3	22	11	50.0
4-8	14	5	35.7
9-12	2	1	50.0

Source: Robinson and Wittebols (1986, 71).

Table 3.3
Class Size and Student Achievement: Studies Clustered by Mathematics Achievement

Grade Level	*Total Number of Studies*	*Studies Favoring Small Class Size*	
		Number	*Percent (%)*
K-3	14	5	35.7
4-8	15	6	40.0
9-12	17	0	0.0

Source: Robinson and Wittebols (1986, 80).

and other variables. The data for the studies are not identical in terms of achievement measures, unit of analysis (classroom or school), or measures of SES; thus, they are not always comparable. Some studies deal solely with class size while others include other aspects of education. In each case, there are differences of opinion regarding the method of analysis and conclusions. The evidence here is presented in the form of tables summarizing selected studies on class size (Tables 1, 2, and 3.1-3.3) and education policy studies (Tables 4-5) so that the reader can evaluate the merits of the conclusions.

Analysis of Controlled Experiments

Looking at the evidence one way, the conclusion seems to be class size does not make a difference, and, therefore, it should not be considered for further funding. Looking another way, the conclusion is that class size does make a difference and should be funded. Looking at the evidence a third way, it is reasonable to conclude that instructional quality and time make the largest difference and should be most heavily funded.

- According to Hanushek (1998, 25): "The economic evidence is clear. There is little reason to believe that smaller class sizes systematically yield higher student achievement. While some studies point in that direction, an almost equal number point in the opposite direction. Moreover, restricting attention to the best of these

Table 4
Production Function Studies

Inputs	*Statistically Significant*	*Statistically Insignificant*
Teacher Characteristics:		
Verbal achievement	12	3
Experience	24	5
SES background	6	1
Gender	1	0
Salary	17	1
Turnover rate	6	3
Employment status	1	0
Job satisfaction	2	1
Teacher personality	1	0
Professional preparation and academic training	18	11
NTE score	3	1
Policy and Administrative Arrangements:		
Class size	10	5
Pupil teacher ratio	13	6
Size of specific class	5	0
Specific staff to pupil ratio	4	0
Paraprofessional assistance for teachers	2	0
Teacher to administrator ratio	2	0
Number of special staff	3	1
Ability groups or tracking practices	6	2
Classroom atmosphere	1	0
Number of days of school	1	0

Source: Betty MacPhail-Wilcox and Richard A. King, "Production Functions Revisited in the Context of Educational Reform," *Journal of Education Finance* 12 (Fall 1986): 203-218, Tables 1-3.

Note: Facilities and fiscal characteristics from original table are not included here.

studies, including those with the most accurate measures of individual class sizes, merely strengthens the overall conclusion."

- According to Krueger (2002, 18): "In sum, all three of these alternatives to Hanushek's weighting scheme produce results that point in the opposite direction of his findings: all three find that smaller class sizes are positively related to performance, and that the pattern of results observed in the 59 studies is unlikely to have arisen by chance."[3]
- According to Robinson and Wittebols (1986, 197): "This research analysis dispels the idea of an 'optimum' class size covering all types of students, in all subject areas and at all grade levels. Students at different grade levels, in different subject areas, and at different levels of personal and academic development require different learning conditions in order for optimum gains in achievement to occur."
- According to MacPhail-Wilcox and King (1986, 220-222): "First, the characteristics of students...may contribute more to the learning process than any purchased resources. Second, teachers' socio-economic status, salary, experience, and verbal abilities are all related to pupils' achievement. Third, professional preparation of teachers is not consistently related to student achievement. Fourth, various indices show particularly strong relationship between student achievement and class size. Finally, levels of expenditures are closely related to student achievement."
- According to Hedges, Laine, and Greenwald (1994, 11): "Taken together, the effect size analyses suggest a pattern of substantially positive effects of global resource inputs (Per Pupil Expenditures) and for teacher experience. The effects of certain resource inputs (teacher salary, administrative inputs, and facilities) are typically positive, but not always. The typical effects of class size (expressed either as pupil/teacher ratio or teacher/pupil ratio) are decidedly mixed."

Each reader must evaluate these materials and statements based on the tables above and/or consult the original documents. The next section attempts to place these materials and conclusions into a larger context.

How Much of a Difference Does Class Size Make on Achievement?

In the previous section, the focus was on the statistical significance of the relationship between class size and achievement. The focus is now on the magnitude and nature of the relationship:

- What is the magnitude of the relationship—the rate of return—or what is commonly called effect size?
- What is the nature of the relationship—does the rate of return change?

These concepts are easily discerned when plotted. The slope of the line indicates the magnitude and the shape of the line indicates a change in the rate of return. There are two basic options for the shape of the line, linear or nonlinear. If linear, there is no change in the rate. If nonlinear, the shape either increases, decreases, or both increases and decreases.

Table 5
Summary of the Production Function Coefficients Utilized in Hedges, Laine, and Greenwald (1994) Analysis

Input Variable	*Number of Estimates*	*Statistically Significant (%)*		*Statistically Insignificant (%)*		
		Positive (%)	*Negative (%)*	*Positive (%)*	*Negative (%)*	*Unknown (Number)*
Per Pupil Expenditure						
Hanushek	65	24	6	46	24	11
Reanalysis	55	24	5	45	25	
Combined significance	35	34	5	37	20	
Effect size estimation	38	27	3	53	18	
Teacher experience						
Hanushek	140	32	8	35	25	15
Reanalysis	131	30	5	40	25	
Combined significance	107	32	7	36	25	
Effect size estimation	57	26	4	46	25	
Teacher education						
Hanushek	113	11	7	41	42	113
Reanalysis	88	11	7	44	38	
Combined significance	68	12	7	51	29	
Effect size estimation	41	10	7	32	51	
Teacher salary						
Hanushek	69	24	9	36	31	24
Reanalysis		21	9	37	33	
Combined significance		23	12	42	23	
Effect size estimation		15	11	37	37	
Teacher-pupil ratio						
Hanushek	152	13	12	32	43	45
Reanalysis		10	13	38	38	
Combined significance		11	13	42	34	
Effect size estimation		9	10	30	51	

Source: Larry V. Hedges, Richard D. Laine, and Rob Greenwald, "Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes," *Educational Researcher* 23 (April 1994): 7, Table 1.

Note: Administrative inputs and facilities were included in the analysis of Hedges et al. (1994), but are not included here.

What Is Class Size?

There are two ways to measure the relationship between the number of pupils and the number of teachers: teacher/pupil ratio; and pupil/teacher ratio. Class size is considered the pupil/teacher ratio. The calculations result in different numerical ratios and have different policy implications. Simply put, school do not have the option of removing students from classroom to achieve a desirable class size, so the only option is to hire more teachers. Therefore, the teacher/pupil ratio is the appropriate policy measure of class size.

What Is Effect Size?

Effect size is the change in achievement measured in standard deviations. In general, effect size is reported under two circumstances. In controlled experiments, effect size is the difference of outcomes between the control and experimental groups measured in standard deviations. In econometric studies, effect size is usually the standard regression coefficient, or the rate of change in the outcome for one standard deviation change in the treatment.

Studies Estimating Magnitude and Shape of the Relationship Between Class Size and Achievement

Below, six studies, four using meta-analysis and two using a controlled experiment approach, are reviewed.

(1) Meta-analysis: Glass and Smith (1978). The research by Glass and Smith was influential in policymaking not because they concluded that class size made a difference in achievement but because they claimed that the influence became larger as classes got smaller. In essence, the effect size became larger as classes became smaller than about 15. To follow is a sampling of statements from other studies attesting to the influence of their proposition.

According to Hanushek (1998):

> The design was heavily influenced by an earlier summary of research by Glass and Smith. That latter study combined the evidence from different experimental studies and suggested that student achievement was roughly consistent across class sizes until the class size got down to approximately 15-to-1. After 15-to-1, reductions in class size appeared to yield significant gains in student performance (p. 26).
>
> Moreover, the original Glass and Smith (1978) analysis itself cast serious doubts on the potential for any improvement in student performance for this policy (p. 37).

According to Mosteller (1995, 115):

> The Tennessee legislators and teachers were also aware of an investigation by Glass and colleagues which reviewed the vast literature on the effects of class size on learning using a special quantitative method called meta-analysis. The results of this investigation suggested that a class size of 15 or fewer would be needed to make a noticeable improvement in classroom performance. At the time of the Glass study, the effect of class size on performance was controversial because many studies in the literature differed in their outcomes.
>
> The new methods used by Glass and his colleagues were not accepted by all professional groups. At the same time, there were ongoing discussions about the lesser cost and possibly equal effectiveness of placing paid teachers' aides in elementary classrooms. Because of the additional expense associated with a reduction in class size for early grades, members of the Tennessee legislature decided that any proposed innovation should be based on solid information and, therefore, authorized a four-year study of class size which would also examine the cost-effectiveness of teachers' aides. The legislature appropriated $3 million in the first year for a study of pupils in kindergarten and then appropriated similar amounts in subsequent years for the project, which carried the acronym STAR (for Student-Teacher Achievement Ratio).

According to Bohrenstedt and Stecher (2002, 22):

> Among the most influential research was Glass and Smith's 1978 meta-analysis of 77 class size reduction studies, which concluded that "large [achievement] advantages [can be expected to occur] when class size is reduced below 20" (Glass and Smith, 1978, p. ii). In a 1982 follow-up report, Glass and associates reiterated the earlier findings and noted that of the more than 100 well-controlled comparisons, 81 percent favored smaller class sizes. They strongly suggested that class sizes needed to be reduced to fewer than 20 pupils for significant results to be observed (Glass et al., 1982).

(2) Meta-analysis: Phelps (2011). (See first article in this issue.) Phelps conducted a reanalysis of Glass and Smith and identified several flaws in assumptions and mathematics. He concluded that the data contained in the meta-analysis indicated a much different relationship between class size and achievement when the contrived methodology was removed. Specifically, Glass and Smith superimposed the squared term into the regression equation to obtain an artificial emphasis on class sizes below 15. Then, to correct for this imposition, they superimposed an entirely different equation on class sizes above 24. Plotting the data without the selection of a "preferred" regression equation,[4] the data showed a complex curve with high points at class sizes of 1, 33, and 64, inconsistent with the original conclusions. The reader is urged to review these findings.

Table 6
Median Regression Coefficients

Input Variable	*Number of Studies*	*Coefficient*
Pupil/teacher ratio		
All studies	45	0.0600
Achievement	22	0.0150
Teacher/pupil ratio		
All studies	24	-0.0010
Achievement	16	0.0176
Teacher education		
All studies	41	-.0200
Achievement	19	-.0300
Teacher experience		
All studies	57	.0700
Achievement	28	.0415
Teacher salary		
All studies	27	.0008
Achievement	12	-.0013
Per pupil expenditure		
All studies	38	.0014
Achievement	26	.0020

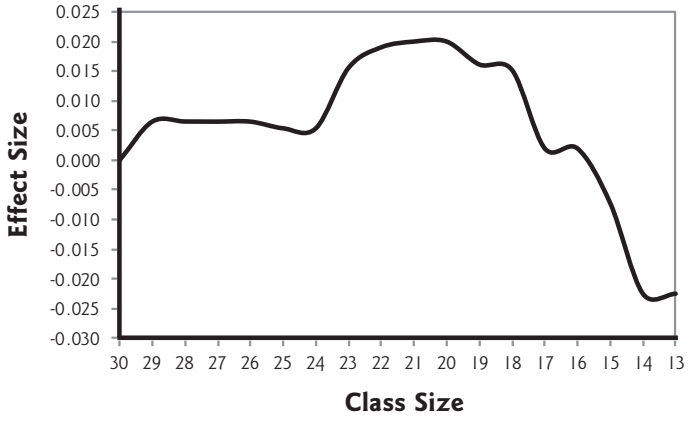
Source: Hedges et al. (1994, 11, Table 4).

(3) Meta-analysis: Hedges, Laine, and Greenwald (1994). Hedges et al. estimated the relationship between several variables and student performance via standard regression coefficients: The amount of change in performance based on the change of an input. The study is a meta-analysis of other studies. Their motivation was to respond to the work of Hanushek (1989) and the implication that money does not matter. (See Table 6.)

Regarding the issue of class size, Hedges et al. (1994, 11) observed: "The typical effects of class size (expressed either as pupil/teacher ratio or teacher/pupil ratio) are decidedly mixed." This is consistent with the Hanushek analysis. Hedges et al. (1994, 11) included a per pupil expenditure variable (PPE) in their analysis and reached the following conclusion: "It [the result] suggests that an increase of PPE by $500 (approximately 10% of the national average) would be associated with a 0.7 standard deviation increase in student outcome."

(4) Meta-analysis: Addonizio and Phelps (2000). Addonizio and Phelps conducted a meta-analysis of four class size studies:

Figure 1
Average Marginal Effect Size across All Subjects and Grades

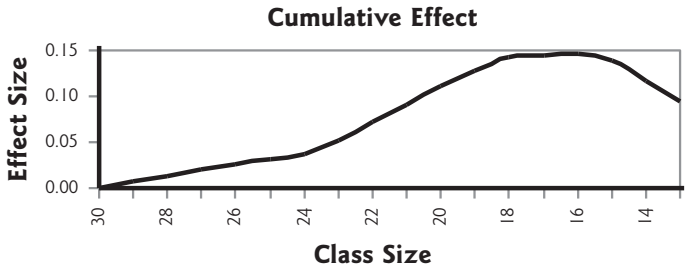


Source: Michael F. Addonizio and James L. Phelps, "Class Size and Student Performance, a Framework for Policy Analysis," *Journal of Education Finance* 26 (Fall 2000): 151, Figure 6.

Tennessee STAR, as reported by Mosteller (1995); Ferguson (1991), Ferguson and Ladd (1996), and Akerhielm (1995). The following is an excerpt from Addonizio and Phelps (2000, 150-154):

> The findings of four studies were summarized in a matrix with the individual outcomes from the studies as the rows, the class size intervals as the columns, and the marginal effects associated with class size changes as the cells. Of course, the cells contain the rates of change in the outcome only for the intervals of change reported in each study; therefore some cells are blank. The estimated effects can be plotted to indicate the general pattern of the effects on measured achievement over the entire range of class sizes. (See Figure 1.)
>
> Again, each cell in the matrix reports the marginal effect over the class size interval. In order to obtain an estimate of the cumulative effect across the range of intervals examined in each study, the average marginal rates of change for each interval are summed. (See Figure 2.)
>
> Finally, the functional relationship depicted in Figure 2 masks the substantial variation in findings across the studies.

Figure 2
Average Cumulative Effect Across Studies

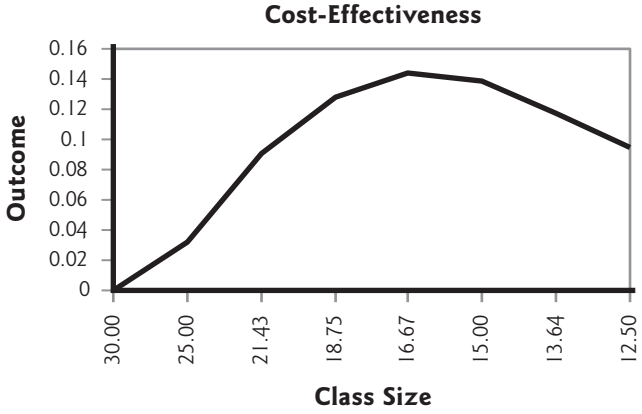


Source: Addonizio and Phelps (Fall 2000, 151, Figure 7).

(Quotation continued)

> These caveats raise questions regarding the appropriateness of combining the results as we have in an attempt to reach a general conclusion about the class size and student achievement relationship. With these caveats in mind, we find that achievement does rise as class size is reduced from about 30 to about 18.
>
> It is one thing to find a statistically significant relationship between class size and student achievement and quite another to determine that investment in smaller classes is a cost-effective strategy. This study has examined the estimated effect sizes of class size reductions from several published studies and will now derive a marginal cost function from these findings.
>
> The class size intervals—30, 29, etc.—provide the starting point for the cost-effectiveness analysis. Of course, the number of teachers necessary to reduce class sizes from 30 to 29 is not the same as reducing the class size from 29 to 28. Each successive incremental reduction in class size requires the hiring of an increasing number of teachers. For example, assuming 150 students in a grade, it would take 5 teachers to produce a class size of 30. By employing an additional teacher (making 6), the class size would then be 25, a reduction of 5. If a second teacher were added, the class size would then be 21.4, a marginal reduction of 3.6 students per classroom. Assuming a cost of $60,000 per teacher, we combine costs and estimated effects to derive a marginal cost curve for improving achievement through class size reductions.
>
> When the relationship between class size and outcomes is adjusted for this cost-effectiveness scale, the relationship looks like (Figure 3):

Figure 3
Cumulative Effect at Various Levels of Resources



Source: Addonizio and Phelps (Fall 2000, 153, Figure 8).

(Quotation continued)

On the basis of our summary of the studies of the generalized relationship between class size and outcomes, the cost-effectiveness analysis indicates a modest gain in outcomes as class size is reduced from 30 about 16, after which the marginal gain falls off.

(5) Controlled Experiment: Mosteller (1995). In 1985, the state of Tennessee started a program to reduce class size in the early grades called STAR (Student Teacher Achievement Ratio). The controlled experiment was structured with two treatment groups and one control group. The control group was the regular-sized classes, and the treatment groups consisted of either smaller classes, or a regular-sized class with an aide. In both treatment groups, achievement was higher than the control group. (See Table 7.)

Mosteller (1995, 125-126) reached this conclusion:

Compelling evidence that smaller classes help, at least in early grades, and that the benefits derived from these smaller classes persist leaves open the possibility that additional or different educational devices could lead to still further gains. For example, applying to small classes the technique of within-class grouping in which the teacher handles each small group separately for short periods could strengthen the educational process (essentially a second-order use of small class size). The point is that small classes can be used jointly with other teaching techniques which may add further gains.

A follow-up study was conducted by Achilles et al. (1993) to assess the long-range benefits of the program. According to Mosteller (1995, 125):

In the Lasting Benefits Study,[5] a continuation of studies evaluated the performance of students from small classes as compared with the performance of students from regular-sized classes or regular-sized classes with an aide after all students had returned to regular-sized classes. The results always favored the students from smaller classes. One year later (1989–90), the effect sizes ranged from 0.11 to 0.16 (n = 4, 230) in the fourth grade, and then, in subsequent years, from 0.17 to 0.34 (n = 4, 639) in the fifth grade, from 0.14 to 0.26 (n = 4, 333) in the sixth grade, and from 0.08 to 0.16 (n = 4, 944) in the seventh grade... Thus, year after year, the students who were originally in smaller classes continued to perform better than the students from regular-sized classes with or without a teacher's aide.[6]

Interestingly, a summary of STAR results appears in *Capstone Report: What We Have Learned about Class Size Reduction in California* (Bohrenstedt and Stecher 2002), indicating the value they placed in the results in hope of a replication.[7]

Project STAR's major findings and those of other research to date include (Finn, 2002):

- Students in small classes performed better at all K–3 grade levels than did students in larger classes.
- Minority and inner city children gained more from reduced classes than their white and nonurban school peers; indeed, the effects were two to three times as great.
- Teacher morale was higher in smaller than in larger classes.

Table 7
Tennessee Class Size Study Summary of Effect Sizes in First Grade

	SAT Reading	*BSF Reading*	*SAT Math*	*BSF Math*
Small class vs. regular-sized class without an aide	.30	.25	.32	.15
Regular-sized classes with an aide compared with regular-sized classes without and aide	.14	.08	.10	.05

Source: Jeremy D. Finn, and Charles M. Achilles, (1990), "Answers and Questions About Class Size: A Statewide Experiment," *American Educational Research Journal* 27 (3): 557-577, Table 5. In Frederick Mosteller, "The Tennessee Study of Class Size in the Early Grades," *Future of Children* 5 (Summer/Fall 1995): 121, Table 2.

- Teachers spent more time on direct instruction and less on classroom management in smaller versus larger classes. Students in smaller classes were more engaged in learning than were students in large classes.
- The earlier and longer the participation in small classes, the greater the effect on achievement.
- Students in small K–3 classes did better academically in grades 4, 6, and 8 than did students in larger K–3 classes.
- The more years students spent in small K–3 classes, the longer-lasting the benefits in later years of schooling.
- Students who had been in small K–3 classes were more likely to graduate from high school, to take college admissions examinations, and, in general, to take courses that prepared them for college than were those who had been in larger K–3 classes. Furthermore, these effects were stronger for minority students, thereby helping close the college preparation gap between African American and white students.

Not everyone reached the same conclusions. Hanushek (1998) argued that the effects in the Tennessee STAR project occurred primarily in kindergarten and first grade and that there was no evidence that additional years of class size reduction contributed incrementally to the effect of small classes in the early years. He acknowledged that the effects were greater for minority and disadvantaged students but then argued, "...the effects appear small relative to costs of programs and alternative policy approaches" (p. 31).

In 1999, Hanushek also took issue with the methodology of the Tennessee STAR project, stating:

While random-assignment experiments have considerable conceptual appeal, the validity and reliability of results depends crucially on a number of design and implementation issues. This paper reviews the major experiment in class size reduction-Tennessee's Project STAR-and puts the results in the context of existing nonexperimental evidence about

(Quotation continued)

class size. The nonexperimental evidence uniformly indicates no consistent improvement in achievement with class size reductions. This evidence comes from very different sources and methodologies, making the consistency of results quite striking. The experimental evidence from the STAR experiment is typically cited as providing strong support of current policy proposals to reduce class size. Detailed review of the evidence, however uncovers a number of important design and implementation issues that suggest considerable uncertainty about the magnitude of any treatment effects. Moreover there is reason to believe that the commonly cited results are biased upwards. Ignoring consideration of the uncertainties and possible biases in the experiment, the results show effects that are limited to very large (and expensive) reductions in kindergarten or possibly first grade class sizes. No support for smaller reductions in class size (i.e., reductions resulting in class sizes greater than 13–17 students) or for reductions in later grades is found in the STAR results (p. 43).

Krueger (2000) countered Hanushek's cost-ineffectiveness argument by showing that there may be significant long-term learning differentials for Tennessee STAR students who were in small versus large classes given that they were more likely to take courses and entrance examinations that rendered them more college ready and, therefore, more job-prepared.

(6) Controlled Experiment: Bohrenstedt and Stecher (1999; 2002). According to Bohrenstedt and Stecher (2002, 4):

A task force assembled by the California Department of Education, called for among other reforms, smaller classes—a move strongly favored not only by the teachers' unions, but also by parents and teachers. California elementary schools had the largest class size in the country—averaging 29 students. Evidence from the Tennessee STAR experiment had shown rather clearly that elementary students in the primary grades did better academically when in small versus larger classes in K–3, and the difference was greatest for inner-city and minority students...A law was passed in July 1996. The law provided districts with $650 per student for each K–3 classroom with 20 or fewer students, providing they first reduced all first grade classes in a school, followed by all second grades and finally by either kindergarten or third grade classes. The cost to the state in the first year was roughly $1 billion dollars and in the current year, roughly $1.6 billion.

In the first report of the CSR Research Consortium (Bohrenstedt and Stecher 1999, 18), there were indications of achievement gain in the smaller classes: "The 'effect size' of the difference between students in smaller and larger classes was nearly 0.1 or one-tenth of a standard deviation. That is equivalent to a 2 to 3 point gain on average in the scale score on the Stanford Achievement Test." The major findings, taken in part from the final CSR report (Bohrenstedt and Stecher 2002, 5-8), are summarized as follows [italics in the original]:

1. *Implementation of CSR occurred rapidly, although it lagged in schools serving minority and low-income students...*
2. *Our analyses of the relationship of CSR to student achievement was inconclusive.* Student achievement has

(Quotation continued)

been increasing since the first administration of the SAT-9 in 1997, but we could find only limited evidence linking these gains to CSR. We found a positive association in 1998 between third-grade class size and SAT-9 scores after controlling for differences in student and school characteristics. However, the size of this CSR effect was small. In the following year, 1998–99, these positive differences persisted when students who had been in reduced size third-grade classes moved to the fourth grade and regular size classes. The spring 1999 SAT-9 results showed that fourth-grade students who had been in reduced size third-grade classes scored higher than those who had not been in such classes. By 2001, CSR implementation was nearly complete, and as a result we could not examine differences in SAT-9 scores between students who were and were not in reduced size classes. Instead, we tracked achievement gains between cohorts of students with incrementally different patterns of CSR exposure to CSR from kindergarten through third grade.

Although both overall exposure to CSR and statewide average test scores increased across cohorts, the magnitude of the changes in test scores did not track with the incremental changes in CSR. Thus, attribution of gains in scores to CSR is not warranted. More refined school-level analyses also failed to find meaningful differences in second- or third-grade scores of students with an additional year of CSR exposure in first grade compared to students who participated only in grades 2 and 3. We could not determine whether our ability to link CSR to achievement was due to weakness of the effect of incremental differences in CSR or to design limitations (or a combination of both). We were also limited in our ability to determine how much of the recent gain in achievement was attributable to CSR and how much was linked to other initiatives.

3. *CSR was associated with declines in teacher qualifications and a more inequitable distribution of credentialed teachers.* Reducing class size required an enormous increase in the number of K–3 teachers in California...To meet the increased demand for teachers, many districts hired teachers without full credentials...Most of the uncredentialed teachers were hired by schools serving the most disadvantaged students, in part because these schools were slower to implement CSR, and more certificated teachers had already been hired elsewhere. In 2000–01, more than one in five K–3 teachers were not fully credentialed in schools with high percentages of low-income, EL, minority, or Hispanic students (primarily large and urban).
4. *CSR had only a modest effect on teacher mobility.* One of the fears was that class-size reduction would result in two types of teacher mobility—teachers from urban schools moving into suburban schools and upper grade elementary teachers moving into K–3. While there was some initial increase, the effect was small and soon disappeared...

(Quotation continued)

5. *CSR implementation did not affect special education identification or placement...*
6. *Students in reduced size third-grade classes received more individual attention, but similar instruction and curriculum.* Compared to teachers with larger classes, teachers of reduced size classes were more likely to say they know what each student knows and can do, that they provide feedback on writing assignments within one day, that they give more individual attention to students, and are able to meet the instructional needs of all students. Teachers in reduced size classes also reported fewer behavior problems and reported that students were more likely to complete the lesson for the day and less likely to be "off task" for more than 5 minutes. But teachers in both reduced and non-reduced size third-grade classes reported spending similar amounts of time and covering similar amounts of curriculum in language arts and in mathematics.
7. *Parents liked reduced size classes.* Based on survey results, parents of third-grade students in reduced size classes rated selected features of their child's education higher than did parents of children in non-reduced size classes. The differences in rating of classroom size were particularly pronounced, with parents of children in reduced size classes reporting satisfaction levels far higher than parents of children in regular size classes. However, parents of children in both reduced and non-reduced size classes expressed equal satisfaction with the qualifications of their children's teachers.
8. *Classroom space and dollars were taken from other programs to support CSR.* Most districts in our statewide sample reported incurring operating costs for CSR that exceeded state payments for it, and these funding problems persisted, or even worsened, in recent years. Districts attempted to overcome budget shortfalls created by CSR by reducing funds for facility maintenance and administrative services. About one-third of such districts also reduced resources for professional development, computer programs, or libraries. To be able to implement the program, many schools reported having to reallocate full-sized classrooms that had been designated for special education back to K-3 classrooms, thereby forcing special education classes to use alternative spaces. CSR implementation also preempted space from such uses as music and arts, athletics, and childcare programs.
9. *In spite of budget shortfalls districts are not projecting CSR cutbacks for 2002-03...*Some [districts] did indicate, however, that cuts to the CSR program were a possibility and would continue to be discussed as their budgets were developed. However, it would be a "last resort" change given the popularity of CSR with parents and teachers.

Effect Size Estimates for Instructional Policy Options

There are few studies estimating the effect size for instructional policy options. Walberg (1984) compiled a comprehensive list of estimated effects in three categories: Student aptitudes; instructional quality and time; and home, peer, class morale, and media. (See Tables 8 and 9.) A class size effect was estimated at .09; however, no class size interval was provided to calculate a rate of change. Walberg (1984, 25) concluded: "Syntheses of educational and psychological research shows that improving the amount and quality of instruction can result in vastly more effective and efficient academic learning. Educators can do even more by also enlisting families as partners, and engaging them directly and indirectly in their efforts."[8]

Table 8
Instructional Quality and Time Effects on Learning

Method	*Effect Size*
Reinforcement	1.17
Acceleration	1.00
Reading training	0.97
Cues and feedback	0.97
Science mastery learning	0.81
Cooperative learning	0.76
Reading experiments	0.60
Personalized instruction	0.57
Adoptive instruction	0.45
Tutoring	0.40
Individualized science	0.35
Higher order questioning	0.34
Diagnostic prescriptive methods	0.33
Individualized instruction	0.32
Individualized mathematics	0.32
New science curricula	0.31
Teachers expectations	0.28
Computer assisted instruction	0.24
Sequenced lessons	0.24
Advance organizers	0.23
New mathematics curricula	0.18
Inquiry biology	0.16
Homogenous grouping	0.10
Class size	0.09
Programmed instruction	-0.03
Mainstreaming	-0.12
Instructional time	0.38

Source: Herbert J. Walberg, "Improving the Productivity of America's Schools," *Educational Leadership* 41 (May 1984): 24, Figure 3.

Table 9
Home, Peer, Class Morale, and Media Effects

Method	*Effect Size*
Graded homework	0.79
Class morale	0.60
Home interventions	0.50
Home environment	0.37
Assigned homework	0.28
Socioeconomic status	0.25
Peer group	0.24
Television	-0.05

Source: Walberg (May 1984, 24, Figure 4).

Effect Size Based on Organizational Effectiveness

Levin (1997) made a case for improving achievement by increasing the effectiveness of school operations. He identified five areas for attention: (1) Commitment to a clear purpose with measurable outcomes; (2) incentives linked to the success of meeting the outcomes; (3) access to useful information for decision-making; (4) flexibility to meet changing conditions; and (5) use of productive technology. Accordingly, efforts towards effectiveness were more likely to improve achievement than increased resource allocations.

Phelps (2009) estimated the effect size of school effectiveness by inspecting the residuals of a production function. The research question was whether schools consistently performed better than their predicted achievement levels when controlled for socioeconomic status (SES), staffing quantity, staff qualifications, and instructional materials. The answer was yes. Over the four-year period, schools consistently either overperformed or underperformed on the achievement expectation. The effect size was measured in terms of the amount of statistical variance explained by averaging the residual. SES explained about 55%, and school and district effectiveness about 27%, supporting Levin's contention.

Other references to this general issue include: (1) In *Cost-Effectiveness and Educational Policy*, Levin and McEwan (2002) addressed many of these issues in great detail; (2) In *Measuring School Performance and Efficiency: Implications for Practice and Research*, Stiefel, Schwartz, Rubenstein, and Zabel (2005) addressed the issues of effectiveness; and (3) In *Making Schools Work: Improving Performance and Controlling Cost*, Hanushek et al. (1994) provided practical alternatives for school improvement.

Some Observations

These questions remain unanswered: (1) Is adding staff a good investment? (2)Will effective instructional and organizational policies produce better achievement results? (3)How should policymakers decide between adding staff or changing instructional and organizational policies?

Hedges et al. (1994, 11) made the following observation:

> It might seem odd that the effect of global resources inputs (PPE) are so clearly positive while the effects for the components are less consistently positive. However, this is not at all contradictory. This pattern of results is consistent with the idea that resources matter, but allocation of resources to a specific area (such as reducing class size or improving facilities) may not be helpful in all situations. That is, local circumstances may determine which resource inputs are most effective, and local authorities utilize discretion in wisely allocating global resources among the areas most in need.

Maybe Hedges et al. are correct: Local circumstances should determine the effective policy options, and uniform statewide or national policies are likely to be ineffective. This might explain why the beneficial effects of statewide policies are difficult to measure and why some schools tend to be associated with higher academic achievement and others are not, even when adjusted for SES and resources.

The Decision-Making Taxonomy

The natural sciences provide many examples where the identification of a unifying structure leads to a new paradigm--a new way to think about the subject, a new way to think about research, and a new way to think about decision-making. To name just a few: the Periodic Table in chemistry; DNA in biology and chemistry; and Gravity, Relativity, and Quantum Mechanics in physics. At the beginning of this article, a decision-making taxonomy was suggested with these underlying questions:

A. Does the research fit into a unifying structure where the evidence and conclusions can be compared and evaluated?
B. Does the research fit into a unifying structure valuable in a decision-making process?

Based on the review of research, below are some observations regarding the decision-making taxonomy.

(1) Professional and public opinion regarding class size. Professional and public opinion matter! The reader is encouraged to reread the Bohrenstedt and Stecher (2002) regarding public opinion. The public is willing to sacrifice other programs to keep lower class sizes in light of budgetary difficulties—even when smaller classes produced no apparent results and at substantial costs. Also reread the section giving credit to the research of Glass and Smith for investing in class size reduction. People believe lower class size works and tend to believe research supporting that position.

Teachers and parents of children in school clearly favor lower class size. Perhaps they see themselves as the beneficiaries of the policy. Legislatures, board members, administrators, and parents without children in school tend to be less enthusiastic, probably because they are more responsible for the funding of a class size policy. Public education is a political entity relying on public opinion. If the public opinion is not accurately informed and changed, moving away from lowering class size to other more cost-effective policies will indeed be difficult. In light of the evidence, a change in opinion is appropriate. A change in the heavy reliance on public opinion by decision makers might also be appropriate. The answers to the underlying questions: A=No; B=No.

(2) A critical analysis of educational research evidence regarding class size: What is statistically significant? Without doubt, the econometric research on class size is mixed. The many meta-analyses show a balance of positive and negative effect signs and a balance of significant and insignificant results. It seems as if the analysis is analogous to a partly filled glass of water: Some see it half-full, and some see it half-empty. Policymakers are in the same position regarding a class size decision; it comes down to personal and public preferences. While the econometric studies were valuable at one time, that time may have passed. More comprehensive research would be more valuable for decision-makers. The answers to the underlying questions: A=No; B=No.

(3) A critical analysis of educational research evidence regarding class size: What is the nature of the relationship? Glass and Smith (1978) contend class size makes a substantial difference in achievement, but only when the classes are smaller then about 15; there, achievement steadily increased as classes become smaller. Phelps, in the first article in this issue, refuted Glass and Smith by identifying shortcomings in their analytical method and by reanalyzing their data with less prejudiced means. The result of the reanalysis shows a pattern of increasing and decreasing benefits to scale, a confusing pattern difficult to interpret or defend. In another meta-analysis, Addonizio and Phelps (2011) found a diminishing returns point where further reductions in class size produced little or no additional gain. This finding was directly the opposite that of Glass and Smith.

There is no clear indication as to the nature of the impact of class size on achievement. In most cases, the assumption is that the relationship is constant—benefits continue for every reduction in class size. But maybe that assumption is incorrect. There are many illustrations where "some" is "good," but "more" either does not add any benefit or could cause harm. It is possible—indeed likely—there are circumstances where there is a benefit threshold, and it is prudent to move to other policy areas when the threshold is reached. The answers to the underlying questions: A=Maybe; B=Maybe.

(4) A critical analysis of educational research evidence regarding class size: What is the magnitude of the relationship? Hedges et al. (1994) found no consistent effect size associated with reducing class size, but found a positive and strong effect size with per pupil expenditures, citing the standard regression coefficients as evidence. Their conclusions were curious:

- The amount of money made a difference, but when spent in the most usual ways, it did not.
- The estimated improvement in achievement for an additional $500 was the same for all schools.
- The estimated improvement in achievement for an additional $500 was the same for every increment of $500, i.e., an increase of $1,500 would produce three time the results of $500.

Here is a thought experiment. Take a hypothetical classroom with 20 pupils and a teacher with a salary of $60,000. The teacher is given $500 per pupil (a total of $10,000) to improve achievement, as suggested by Hedges et al. However, the condition is that achievement must improve by .7 standard deviations or the teacher will forfeit $10,000 of their salary. To make the conditions fairer, the teacher selects his or her students, either high-achieving or average-achieving.[9] What are the chances of the teacher being successful? Would a reasonable teacher accept these conditions?

Hedges et al.'s conclusion regarding the achievement result of a $500 investment is a reasonable interpretation of the standard partial regression coefficient, but these findings are in conflict with the conclusion stated earlier: Benefits accrue based on individual school decisions. The implication of the Hedges et al.'s proposition is that all schools will get the same results with the same additional expenditures, but this is not the case. The regression line is not actually a line; it is a three-dimensional distribution with the average of the distribution being the regression line; [10] that is to say, at any expenditure level, half of the schools will do better than what the line predicts and, half will not do as well. To express it another way, some schools are more effective than others in how they spend money. Economists call this efficiency.[11] It stands to reason if the ineffective schools spend the new money in the old way, there is little chance the predicted achievement gain will be realized, but if they spend the new money in a more effective way, the gains could be larger. This scenario raises an unusual dilemma. What if the ineffective schools would have spent the previous money more effectively? Surely their achievement scores would be higher. With this interpretation of the regression statistics, the logical answer is not to spend more money but to spend the existing money more wisely. Hedges et al.'s own analysis demonstrated the areas where schools spend money with no achievement benefit--teacher education, teacher salary, and administrative inputs. A case could be made that additional money could be helpful in making the effective changes in the school instructional programs or in the operations of the organization. Economists call these "opportunity costs." As suggested by Levin (1997) and measured by Phelps (2009), these opportunities are likely to be substantially larger than what would accrue with more resources. The answers to the underlying questions: A=Likely; B=Likely.

There is another consideration in the Hedge et al.'s interpretation. It is unlikely that the top-performing schools will accrue the same benefit as the lowest-performing schools with the same dollar amount and the same degree of effectiveness—there is a performance ceiling effect. Because there is an upper limit to achievement tests, high performing schools have larger numbers of students near or at the test ceiling; they have no room to improve. Another example of a ceiling is teacher experience. The interpretation of standard partial regression coefficients is that for every additional year of experience achievement will increase by the same amount—only if the teachers do not exhibit the same behavior each year. Clearly experience matters because as new teachers gain experience they change their behavior, but after a period of time, say five years, the changes are minimal. There is a behavior ceiling unless there is a change in the operations of the school or the instructional program. It is doubtful whether a prudent teacher, knowing the other interpretations of the statistics, would accept the thought experiment challenge. The moral: Don't always bet on the standard partial regression statistics! The answers to the underlying questions: A=Likely; B=Likely.

(5) A critical analysis of educational research evidence regarding class size: What do controlled experiments say about the magnitude of the relationship? The analysis of the Tennessee controlled experiment found positive and substantial benefits with effect size around a standard deviation, or effect size, of .25 for the smaller classes and .09 for regular classes with an aide (Achilles et al., 1993). The results for mathematics were about .04 lower than for

reading. On the other hand, the analysis of the California controlled experiment found no achievement gain attributable to the reduction in class size (Bohrenstedt and Stecher, 2002), although there was an effect size of about .10 reported in an early analysis (Bohrenstedt and Stecher, 1999). There were not enough instructional or organizational data collected to explain why the results might be different in these situations. Surely, the different results were not due to the difference in location or time period. There must have been different circumstances. Were there differences in the instructional programs or the operations of the organizations?

While the controlled experiments estimated effect size, it is not the same measure as reported in the econometric studies. The experiments reported the effect difference between treatment and control groups while the econometric studies reported an effect rate of change, or a change in achievement for a given change in class size. The answers to the underlying questions: A=Unclear; B=Unclear.

(6) A critical analysis of educational research evidence regarding class size: What is the cost-benefit relationship? There is no disputing the fact that lowering class size is costly. Most of the econometric analyses do not focus on this point. Levin (1997) and Phelps (2009) demonstrated the concepts, methods, and benefits of cost-effectiveness analysis. The answers to the underlying questions: A=Likely; B=Likely.

(7) A critical analysis of educational research evidence: What is the magnitude of the relationship between achievement and instructional policy options? Walberg (1984 suggested that instructional and time policies have a major influence on achievement.His estimates of effect size raised several puzzling questions:

- Because the effect size estimates were substantially larger than those of class size, why is there so much emphasis on lowering class size?
- If the instructional and time benefits were so large, why don't schools implement these policies?
- If schools implemented the instructional and time policies and they were of the suggested magnitude, why aren't the results apparent in the improvement of overall achievement in the U.S.?
- Is it possible the effect sizes were overestimated?

There is an underlying impression that each of the instructional and time policy options operate independently—substantial achievement gains will be realized with each action taken—because the policy options are unique and additive. That impression is most likely false. More likely, there is a commonality among these instructional policy options suggesting they work together rather than separately and, as a result, there is a ceiling to their overall contribution. Actually, this notion is inherent in the nature of achievement testing and in the regression formulation. There is a ceiling to achievement tests, the perfect score. No matter the effect sizes, they cannot add up to perfect scores for all students because the tests are made to identify differences among students. Without variance in the tests, they would serve no useful purpose. There is a test ceiling with built-in variance. Regarding regression, if the instructional and time policy variables are correlated, and they surely are, they share a common variance. As a result, as variables are added, their contribution to the total explanation is increasingly smaller—the basis of stepwise regression. The answers to the underlying questions: A=Likely; B=Likely.

(8) A critical analysis of educational research evidence: What is the magnitude of the relationship between achievement and organizational policy options? Levin (1997) suggested that effective operation of the school has more to do with improving achievement than the allocation of resources. Phelps (2009), following up on the Levin proposition, estimated the effect size of instructional and organization effectiveness to be substantially higher than that for the allocation of resources. Their work supports the idea that effective utilization of the resources is more important than the amount of the resources, counter to the Hedges et al. (1994) proposition. The implications are enormous. There are many ineffective schools due to their operations, not due to the level of resources or SES. Conversely, there are many effective schools due to their operations, not due to the level of resources or SES. This important conclusion is repeated: The effect size attributable to effectiveness is large, substantially larger than what can be attributed to class size or any other resource policy. In other words, the success of implementing any resource policy is more dependent on the level of effectiveness than the policy itself.

Is it possible to determine what effective schools are doing and provide the knowledge to the others? Unfortunately, there is little research as to the reasons for the effectiveness. However, it is possible to include the concept of effectiveness in the policy analysis process. The answers to the underlying questions: A=Likely; B=Likely.

(9) A decision-making process including: Establishing a set of clearly stated goals; identifying a set of possible policy options to achieve the goals; clearly stating the assumptions why each of the policy option would achieve the goals; and evaluating each of the policy options to select the best alternative. If the above statement reflects the highest category on the suggested decision-making taxonomy, then existing research is scant. Without a clear statement of the underlying assumptions regarding the potential benefits of the competing alternatives and a practical decision-making model, what remains are personal preferences. These preferences morph, as Hedges 1994) suggested, into local discretion. In many cases, this process clearly works, as measured by the results; but, in other cases, it clearly does not, and a closer look at the decision-making process seems warranted.

The difference between level one and level three of the decision-making taxonomy, and the reasons why level one is the most common, is captured in the following quote from Schrage (1991, 305):

> The advantage and perhaps the major motivation for using "seat-of-the-pants" decision making is that it obscures the assumptions made in arriving at a decision. If no one knows the assumptions upon which you based your decisions, then even though they may be uneasy with the decision they will have a difficult time criticizing your assumptions or decisions.

What is missing in the research review is an integrated and comprehensive paradigm capable of accommodating the seemingly unrelated research and dissimilar numerical estimates into a unified structure conducive to policy analysis and decision-making.

Kuhn (1970), author of *The Structure of Scientific Revolutions*, is noted for his thoughts regarding paradigms. He set two essential characteristics: The work was "sufficiently unprecedented," from competing modes of research, and "sufficiently open-ended with all

sorts of problems to resolve" (p. 10). He continued to describe the characteristics as including theory, mathematical laws, applications, instrumentation, and rules for future research. Later, Kuhn (1970, 15) made an observation which appears to summarize the previously reviewed research:

> In the absence of a paradigm or some candidate for paradigm, all of the facts that could possibly pertain to the development of a given science are likely to seem equally relevant. As a result, early fact-gathering is a far more nearly random activity than the one that subsequent scientific development makes familiar. Furthermore, in the absence of a reason for seeking some particular form of more recondite information, early fact-gathering is usually restricted to the wealth of data that lie ready to hand.

The nine points identified above are a modest attempt at building a conceptual base for such a policy analysis paradigm. The following articles in this issue will combine the various estimates of effect sizes into a coherent structure (theory and laws); build a rationale (theory) and analytical method (laws) to accommodate the ceiling and effectiveness effects; and demonstrate an integrated and comprehensive policy analysis paradigm (instrumentation and application).

References

Achilles, C.M., B.A. Nye, J.B. Zaharias, and B.D. Fulton. "The Lasting Benefits Study (LBS) in Grades 4 and 5 (1990–1991): A Legacy from Tennessee's Four-year (K–3) Class-size Study (1985–1989)." Project STAR. Paper presented at the North Carolina Association for Research in Education, Greensboro, North Carolina, January 14, 1993.

Addonizio, Michael F., and James L. Phelps. "Class Size and Student Performance: A Framework for Policy Analysis." *Journal of Education Finance* 26 (Fall 2000): 135-156.

Akerhielm, Karen. "Does Class Size Matter?" *Economics of Education Review* 14 (June 1995): 229-241.

Barnett, W. Steven. "Benefits of Compensatory Preschool Education," *Journal of Human Resources* 27 (Spring 1992): 279-312.

Bloom, Benjamin S. "The Search for Methods of Groups Instruction and as Effective as One-to-One Tutoring." *Educational Leadership* 4 (May 1984): 4-17.

Bohrenstedt, George W., and Brian M. Stecher, eds. *Class Size Reduction in California: Early Evaluation Findings, 1996-98*. Sacramento, CA: CSR Research Consortium, California Department of Education, June 1999. http://www.classize.org/techreport/index.htm.

Bohrenstedt, George W., and Brian M. Stecher, eds. *Capstone Report: What We Have Learned about Class Size Reduction in California*. Sacramento, CA: CSR Research Consortium, California Department of Education, August 2002. http://www.classize.org/techreport/index-02.htm.

Ferguson, Ronald F. "Paying for Public Education: New Evidence on How and Why Money Matters." *Harvard Journal of Legislation* 28 (Summer 1991): 465–498.

Ferguson, Ronald F., and Helen F. Ladd. "How and Why Money Matters: An Analysis of Alabama Schools." In *Holding Schools Accountable: Performance-Based Reform in Education*, edited by Helen F. Ladd, 265-298. Washington, DC: The Brookings Institution, 1996.

Finn, Jeremy D. "Class-Size Reduction in Grades K-3." In *School Reform Proposals: The Research Evidence*, edited by Alex Molnar, 27-48. Greenwich, CT: Information Age Publishing, 2002.

Glass, Gene V., and Mary Lee Smith. *Meta-analysis of Research on the Relationship of Class-size and Achievement*. San Francisco, CA: Far West Laboratory for Educational Research and Development, 1978.

Hanushek, Eric A. "Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects." *Educational Evaluation and Policy Analysis* 21 (Summer 1999): 143-163.

_____. "The Evidence on Class Size." Occasional paper 98-1. Rochester, NY: University of Rochester, Wallis Institute of Political Economy, February 1998. http://edpro.stanford.edu/Hanushek/files_det.asp?FileId=114.

_____. "The Impact of Differential Expenditures on School Performance." *Educational Researcher* 18 (May 1989): 45-65.

Hanushek, Eric, A., and Dongwook Kim. "Schooling, Labor Force Quality, and Economic Growth." NBER Working Paper No. 5399. Cambridge, MA: National Bureau of Economic Research, 1995.

Hanushek, Eric A., with others. *Making Schools Work: Improving Performance and Controlling Cost*. Washington, DC: The Brookings Institution, 1994.

Hedges, Larry V., Richard D. Laine, and Rob Greenwald. "Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes." *Educational Researcher* 23 (April 1994): 5-14.

Kuhn, Thomas S. *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press, 1970.

Krueger, Alan B. "Understanding the Magnitude and Effect of Class Size on Student Achievement." In *The Class Size Debate*, edited by Lawrence Mishel and Richard Rothstein, 7-35. Washington DC: Economic Policy Institute, 2002. http://edpro.stanford.edu/hanushek/admin/pages/files/uploads/classsizedebate.full%20volume.pdf.

Levin, Henry M. "Cost-Effectiveness and Educational Policy." *Educational Evaluation and Policy Analysis* 10 (Spring 1988): 51-61.

_____. "Raising School Productivity: An X-Efficiency Approach." *Economics of Education Review* 16 (June 1997): 303-311.

Levin, Henry M., and Patrick J. McEwan, eds. *Cost-Effectiveness and Educational Policy*. Larchmont, NY: Eye on Education, 2002.

MacPhail-Wilcox, Betty, and Richard A. King. "Production Functions Revisited in the Context of Educational Reform." *Journal of Education Finance* 12 (Fall 1986): 191-222.

Marzano, Robert J., Barbara B. Gaddy, and Ceri Dean. *What Works in Classroom Instruction*. Aurora, CO: Mid-Continent Research for Education and Learning (McREL), August 2000.

Mosteller, Frederick. "The Tennessee Study of Class Size in the Early Grades." *Future of Children* 5 (Summer/Fall 1995): 113-127. http://www.princeton.edu/futureofchildren/publications/docs/05_02_08.pdf.

Nye, B.A., J.B. Zaharias, B.D. Fulton, et al. *The Lasting Benefits Study: A Continuing Analysis of the Effect of Small Class Size in Kindergarten through Third Grade on Student Achievement Test Scores in Subsequent Grade Levels*. Seventh grade technical report. Nashville, TN: Center of Excellence for Research in Basic Skills, Tennessee State University, 1994. Cited in Frederick Mosteller, "The Tennessee Study of Class Size in the Early Grades." *Future of Children* 5 (Summer/Fall 1995): 113-127.

Phelps, James L. "Measuring and Reporting School and District Effectiveness." *Educational Considerations* 36 (Spring 2009): 40-52.

Robinson, Glen E., and J. H. Wittebols. *Class Size Research: A Related Cluster Analysis of Decision Making*. Arlington, VA: Educational Research Services, Inc., 1986.

Schrage, Linus E. *Lindo: An Optimization Modeling System*. San Francisco, CA: Scientific Press, 1991.

Stiefel, Leanna, Amy Ellen Schwartz, Ross Rubenstein, and Jeffrey Zabel, ed. *Measuring School Performance and Efficiency: Implications for Practice and Research*. Larchmont, NY: Eye on Education, 2005.

Walberg, Herbert J. "Improving the Productivity of America's Schools." *Educational Leadership* 41 (May 1984): 19-27.

Woessmann, Ludger, and Martin R. West. "Class-Size Effects in School Systems Around the World: Evidence from Between-Grade Variation in TIMSS." Cambridge, MA: Harvard University, Education Policy and Governance, 2002. http://www.eric.ed.gov/PDFS/ED467039.pdf.

Endnotes

[1] Wherever possible, the original material is presented so that the reader can judge the significance firsthand.

[2] For a complete list of Hanushek's publications, see http://edpro.stanford.edu/Hanushek/content.asp?contentId=81.

[3] Krueger (2002, 16) went on the estimate the amount of variance explained by class size to be 0.08.

[4] This analytical method was used by Addonizio and Phelps (2000) and is described later in this section.

[5] C.M. Achilles, B.A. Nye, J.B. Zaharias, and B.D. Fulton, "The Lasting Benefits Study (LBS) in Grades 4 and 5 (1990–1991): A Legacy from Tennessee's Four-Year (K–3) Class-Size Study (1985–1989), Project STAR, paper presented at the North Carolina Association for Research in Education. Greensboro, North Carolina, January 14, 1993.

[6] B.A. Nye, J.B. Zaharias, and B.D. Fulton, et al. *The Lasting Benefits Study: A continuing analysis of the effect of small class size in kindergarten through third grade on student achievement test scores in subsequent grade levels*. Seventh grade technical report. Nashville, TN: Center of Excellence for Research in Basic Skills, Tennessee State University, 1994.

[7] Finn, a coauthor on Tennessee STAR project publications, served on the CSR advisory panel (http://www.classize.org/advpanel/index.htm), so it is reasonable to assume he participated in preparing this summary.

[8] Following are some other studies regarding instructional effect sizes: (1) In *What Works in Classroom Instruction*, Marzano, Gaddy, and Dean (2000) provided effect size estimates similar to those of Walberg (1984), but provided a full description of the instructional conditions; (2) In "The Search for Methods of Groups Instruction and as Effective as One-to-One Tutoring," Bloom (1984) provided the effect sizes for instructional methods of mastery learning and tutorial instruction all with a consistent class size of 1 to 30; (3) In "Benefits of Compensatory Preschool Education," Barnett (1992, 297) estimated the effect size of preschool programs at .75; and (4) In *Capstone Report: What We Have Learned about Class Size Reduction in California*, Bohrenstedt, George W., and Brian M. Stecher (2002) included references to instructional policy options other than class size reduction.

[9] Starting at the average, the 50th percentile, a .7 improvement would raise the standing to the 75th; starting at the 75th, the improvement would be to the 95th; starting at the 95th, the improvement would be to the 99th. As the starting point gets higher, the percentile gains gets smaller.

[10] The standard error of estimate is the parameter of the three-dimensional distribution.

[11] The efficiency portion of the residual is separated from the random error portion by averaging over time the residual for each observation. In econometrics, this is known as the fixed effect.

A Practical Method of Policy Analysis by Estimating Effect Size

James L. Phelps

The previous articles on class size and other productivity research paint a complex and confusing picture of the relationship between policy variables and student achievement. Missing is a conceptual scheme capable of combining the seemingly unrelated research and dissimilar estimates of effect size into a unified structure for policy analysis and decision making. This article builds a rationale for a unifying structure and consistent method of estimating effect size.

Forrester (1980), in his work on system dynamics, offers pertinent ideas. He stressed the importance of constructing a comprehensive operating structure to better understand an organization's complexity and its behavior in response to policies. By structure, he meant all the diverse elements of the organization, including their specific responsibilities and, most importantly how the elements related to one another in some quantifiable manner. Within the identified operating structure, policy decisions were made to directly influence changes in behavior in specific elements of the organization. Those same policies also indirectly influenced other elements of the organization because the elements were interrelated. Quantifying these elements and their interrelationships within a unified scheme is essential to the workings of system dynamics. This model relies on a set of parameters to simulate organizational behavior in response to various policy options. The purpose of the model is to predict how policy changes will influence organizational behavior which, in turn, will achieve the desired outcomes.

Another representation of the organization is what economists call a production function. The outcomes (outputs) of the organization are the byproducts of the resources (inputs) and the processes used to convert the resources into outcomes. Using this framework, the educational outcomes are achievement measures; the resources are services and materials purchased, e.g., staffing; and the processes include the curriculum, instructional program, and home activities, for example. In most production function studies, however, little attention is paid to the process variables largely because of the lack of data and a meaningful method of assimilation. When interpreting the results, primary attention is directed to the linear weights, or regression coefficients. Less attention is paid to the statistics describing the explained variance (R^2) and the residual. These statistics provide a different approach to a unified structure and method of estimating effect size. The main purpose of the production function is to estimate the parameters of a small set of relationships and make probability inferences. Most econometric studies focus on class size or some other narrow aspect of education rather than the entirety of school activities. As a result, econometrics has substantial limitations in simulating organizational behavior for multiple goals and policy options.

A desirable paradigm would combine features from both system dynamics and econometric modeling. A semantic clarification is in order. Here, I am referring to a paradigm as a model, and a model as a hypothetical formulation used in analyzing or explaining something. In the context of this article, the paradigm is the formulation of a unified school structure including what Kuhn (1970) labeled theory, laws, application, and instrumentation. The model is the mathematical representation of the paradigm, or the laws, application, and instrumentation components of the paradigm. Based on these concepts, the immediate task is to identify the resource and process elements of the educational organization and quantify their relationships with the outcomes, all under some unifying scheme or structure—in other words a paradigm.

This article develops a policy analysis paradigm by combining the various estimates of effect sizes into a coherent structure with a consistent method of measurement; and building a rational and analytical method to accommodate the effect ceiling and effectiveness components. The final product is a suggested analytic structure, a list of characteristics associated with the method of measuring effect size, and a list of assumptions underlying the policy analysis paradigm. Finally, there is a compilation of estimated effect sizes.

What makes this paradigm "sufficiently unprecedented," to use Kuhn's phrase, is the method of estimating effect size permitting the principles of system dynamics to be incorporated into a method of policy analysis. The effect sizes, when coupled with the incremental cost of the policy options, provide policymakers with a model to evaluate the potential achievement gains based on various combinations of alternatives (Kuhn's application and instrumentation). This final stage of the paradigm addresses three overarching questions:

- Under what circumstances might lowering class size be effective?
- What are the competing resource and process policies for improving achievement?
- How do policymakers decide what is the most effective and efficient course to follow?

The first section in this article reviews the conceptual issues related to the relationship between class size and achievement, as follows: Measurement of the concentration of teachers and students; collinearity among the data variables; influence of socioeconomic status (SES) as an intervening variable; and modeling the relationship between achievement and policy options. Section two provides estimates of effect size from a Minnesota data set, utilizing different statistical methods to illustrate the various methods available to measure the magnitude of effect size. It highlights the difficulties in measuring effect size and demonstrates a method to

James L. Phelps holds a Ph.D. from the University of Michigan in Educational Administration. He served as Special Assistant to Governor William Milliken of Michigan and Deputy Superintendent in the Michigan Department of Education. Active in the American Education Finance Association, he served on the Board of Directors and as President. Since retirement, he spends a great deal of time devoted to music, composing and arranging, playing string bass in orchestras and chamber groups, as well as singing in two choirs. He resides with his wife, Julie, in East Lansing, Michigan.

place the various estimates into a unified structure. These estimates are compared with those from the studies reviewed in the previous article. Section three summarizes the material presented and states the assumptions guiding a policy analysis model.

Conceptual Issues

Measurement of the Concentration of Teachers and Students

The method of measuring the concentration of teachers and students has cost implications as demonstrated by this example: The additional cost of reducing the class size from 20 to 19. This raises a concept from physics known as the quantum jump, or the energy required for an electron to jump from one energy state to another. (The energy comes only in well-defined packets. Such is the case with class size.) If there are 60 students in a particular grade, then class size is determined by the number of teachers assigned to that grade. The number of teachers is the quantum number, not the number of students.[1] With 1 teacher, the class size is 60; with 2, the class size is 30; with 3, it is 20; and, with 4, it is 15. In other words, there is no possible way of reducing class size from 20 to 19. In order to lower the class size below 20, the only policy alternative is to add one additional teacher and pay the costs to reduce the class size from 20 to 15. Therefore, the appropriate policy-oriented class size measure is the teacher/pupil ratio.

Collinearity among Explanatory Variables

There is no perfect way to measure effect size. First, there is always a degree of measurement error. Second, in most cases, explanatory variables are intercorrelated. For example, in the case of two explanatory variables, the influence (proportion of variance explained, or R^2) is divided into segments: The unique influence of each variable and the common influence among the variables. There is no unequivocal way to partition the common influence into the unique influence of both variables. The regression process attributes the common influence to the variable with the highest correlation with the achievement variable, most likely SES. Therefore, the variable of policy interest, the teacher/pupil ratio, is allocated the remaining portion of the explained variance and, as a result, a lower weighting. When there are two variables, the compromise is to estimate the maximum effect size (with the common variance) and minimum effect size (without the common variance) for the policy variable and select the appropriate value on other grounds. This same principle applies to the many instructional variables identified by Walberg (1984)[2] and explains why his estimated effect sizes could not be added—they were correlated! When there are more than two variables, it is desirable to combine the effect sizes into a cluster, or factor, containing all the unique and common variance (Phelps, 2009).

Influence of Socioeconomic Status (SES) as an Intervening Variable

Over the years, federal and state governments have provided additional funds to low performing schools. These are determined in a number of ways, usually by achievement scores or SES. Schools receiving these funds often reduce their class size. As a result, it is likely that low-performing schools have lower class sizes. To adjust for this situation, a measure of SES in the analysis is critical. The inclusion of this intervening variable could materially change the magnitude of the relationship between achievement and the policy variable.[3]

Modeling the Relationship between Achievement and Policy Options

Regression is a statistical model to estimate the relationship between policy variables and achievement, but it has limitations pertaining to policy analysis. Because there can be but one regression equation, multiple achievement measures and variables with differing costs are not accommodated. There are other mathematical models addressing these shortcomings which are more helpful in evaluating policy alternatives. These models depend on simultaneous equations and nonlinear relationships between the outcome and the explanatory variables. There are substantial differences between nonlinear and linear models.

Effect size for linear relationships: Constant slope. Linear regression coefficients are the most frequent measure of effect size. The maximum effect size is estimated by regressing only the target variable with the achievement outcome either by the "b" weight or the standard regression coefficient expressed as Beta (β). The standard regression coefficient is more practical because it easily compares variables measured in differing metrics. SES could well be associated with class size, so it should be included as an intervening variable in the multiple regression equation to estimate the minimum.

Effect size for nonlinear relationships: Changing slope. It is highly unlikely that any policy variable will have a consistent, increasing or decreasing slope. Slight variations in the slope can be estimated by adding a squared term to the regression equation.[4] This does not provide either a theoretical or practical solution. There is, however, a theoretical sound and practical solution. This solution utilizes the amount of variance explained by the explanatory or policy variable in question, or the R^2.[5]

The R^2, when interpreted as the cumulative area under the normal curve, produces an S-shaped curve asymptotic at the top (maximum of 100th percentile) and bottom (minimum of zero percentile). If the R^2 is .5, then the S-shaped curve is reduced to the 75th percentile at the top and the 25th percentile at the bottom. As the R^2 approaches zero, the S-shaped curve approaches a line at the 50th percentile.

Mathematical reason for the nonlinear relationship. The difference between the linear and nonlinear interpretations can be demonstrated with a thought experiment using standard regression coefficients (β's). The regression equation states that the predicted outcome (measured in Z-scores) is equal to the sum of the β's times their respective Z-scores (and a percentile ranking can be calculated from any β and Z-score combination):

$$y_{(z)} = \beta_1 Z_1 + \beta_2 Z_2 + \ldots \beta_n Z_n$$

The following calculations are for two hypothetical situations: (1) all Z-scores equal 1 (Z=1); and (2) all Z-scores equal 3 (Z=3). The variables are, SES, teacher/pupil ratio, instruction, and effectiveness. For each β*Z term in the equation, a percentile is calculated to measure the contribution to the overall change in performance. Assuming the starting point is the mean, the percentiles greater than .50 are calculated to determine the predicted gain. The percentile gains for the individual variables are then summed as indicated by the equation. (See Table 1.)

When each of the four variables is increased by 1-Z-score (from zero to 1), the increased percentile standing for all variables is .4236, or from .50 to .9236. When each variable is increased 3-Z-

Table 1
Calculation of Percentiles from Beta (β)

Variables	*β*Z (Z=1)*	*Percentile*	*Percentile >.50*	*β*Z (Z=3)*	*Percentile*	*Percentile >.50*
SES	0.8457	0.8011	0.3011	2.5371	0.9944	0.4944
Teacher/Pupil Ratio	0.0677	0.5270	0.0270	0.2031	0.5805	0.0805
Instruction	0.1200	0.5478	0.0478	0.3600	0.6406	0.1406
Effectiveness	0.1200	0.5478	0.0478	0.3600	0.6406	0.1406
Sum			0.4236			0.8560

Table 2
Calculation of Percentiles from R^2

Variables	*R^2*	*$R^2/2$*	*Z-Score*		
			- infinity	*Z=0*	*+ infinity*
SES	0.6827	0.3414	0.1587	0.5	0.8414
Teacher-Pupil Ratio	0.0280	0.0140	0.4860	0.5	0.5140
Instruction	0.0600	0.0300	0.4700	0.5	0.5300
Effectiveness	0.1400	0.0700	0.4300	0.5	0.5700
Subtotal	0.9107	0.4554	0.0447	0.5	0.9554
Error	0.0893	0.0447	0.4554	0.5	0.5447
Total	1.0000	0.5000	0.0000	0.5	1.0000

scores (from zero to three), the increased percentile standing is .8560. Because the starting point was the mean (.50), the increase brings the total to the impossible 1.356th percentile! Clearly, not all variables can be increased simultaneously. The β weights are partial regression coefficients and assume that all other variables stay fixed.

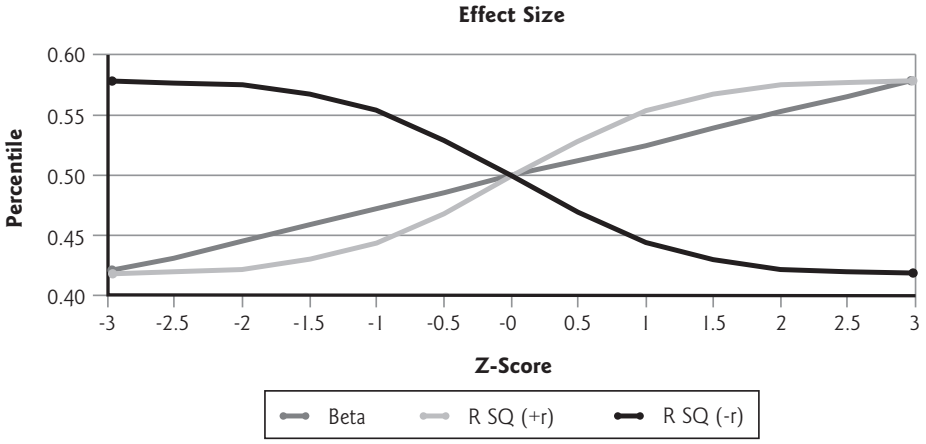
A second example uses the proportion of explained variance, or R^2, as the measure of effect size. To obtain the R^2, β is multiplied by the correlation coefficient: $R^2 = \beta_i r_i$. The R^2 has four advantageous properties. First, the area under the normal curve is by definition equal to 1, so any point on the distribution can be defined as a percentile—the percent of observation below the point. Second, the highest point on the distribution is the 100th percentile and the lowest point is zero percentile. Third, the R^2 is the ratio between the outcome distribution and the explanatory distribution, so a percentile contribution to the outcome can be determined for any point on the explanatory distribution. Fourth, the mean (Z=0) on the explanatory variable will predict the mean of the outcome variable. Table 2 illustrates the percentile range (Z-score of +/- infinity) for each explanatory variable. One-half of the R^2 contribution is above the mean and one-half below. The R^2 values are listed with the minimum and maximum percentile levels. The contribution of the explanatory variables totals .4554 percentile points, ranging from .0447 to .9554.

Because the maximum R^2, including the error, for the variables is 1.00, no combination of variables, regardless of the Z-score can ever be higher than the 100th percentile or lower than zero percentile. In this case, there is no partial or fixed restriction as is the case with the regression β's. All variables are free to vary from the highest to the lowest Z-scores, accommodating the ceiling effect.

Figure 1 illustrates these different interpretations of effect size. The straight line represents the Beta coefficient between the extremes of Z-scores from zero to 3, but with all other variables fixed. The percentile ranking will continue to increase as the Z-score increases. The R^2 curve, the cumulative normal curve, is also between the extreme Z-scores, but with all other variables free to move. In contrast, the curve approaches a ceiling. The R^2 of any variable will have a negative sign if the regression coefficient is negative, as illustrated in Figure 1. The graph clearly depicts the difference between the unbounded character of the Beta coefficient and the ceiling character of the R^2.

Policy analysis differences between linear and nonlinear relationships. If a linear relationship is assumed with the β weight as the measure of effect size:

Figure 1
Representation of Beta (β) Weights and R^2 as Measures of Effect Size



Note: This graph is not to scale: Beta does not equal R^2.

- There can be only one best cost-effective policy, i.e., the variable with the largest standard regression coefficient (β) adjusted for cost.
- There is no reason to adopt anything but the most cost-effective policy option.
- The most cost-effective policy applies equally to all schools.
- There is never a point of diminishing returns.
- The linear relationships do not allow for an optimization process; i.e., finding the best combination of variables and costs to maximize the goals.
- Linear relationships are not an accurate representation of achievement production.

If a nonlinear relationship is assumed with R^2 as the measure of effect size and the residual as the measure of school effectiveness:[6]

- There is no one best cost-effective policy.
- The potential benefits will depend on the unique history of each school, i.e., their existing levels on all the policy variables, requiring unique policies for each school.
- When the benefit of a policy has reached a point of diminishing returns (high point on the S-shaped curve), a different policy with greater potential then becomes the preferred option.
- Nonlinear relationships are a more accurate representation of achievement production.

Recall the dilemma of Hedges, Laine, and Greenwald (1994)[7] as identified in the previous article; that is, spending money would improve achievement in every school even though no specific object for the funds was identified. Likewise, Glass and Smith (1978)[8] advocated lowering class size until there was one teacher for every pupil in order to achieve the maximum potential achievement. The list of instructional programs by Walberg also gave the same impression. In sum, if more funds, lower class size, and more instructional programs were provided, all schools would have unlimited success in raising achievement scores. No attention was paid to the ceiling imposed by achievement tests. No attention was paid to the uniqueness of every school setting. No attention was paid to the effective use of the resources or the quality of the instructional programs. Conclusions were based on the same mathematical model, the boundless regression line, which does not represent the realities of school operations.

If a different mathematical model is employed, one based on the statistical variance around the line, an entirely different notion emerges. Resources and instructional programs do make a difference, but the size of the difference is limited by the achievement test ceiling. The magnitude of these differences depends on the unique circumstances of each school, in contrast to a one policy fits all approach. While resources and instructional programs are important, so is their effective implementation. Because the variance interpretation of the regression statistics more accurately represents the realities of school operations, it is the basis of estimating effect size and simulating organizational behavior.

Estimating Effect Size: Illustrations from the Minnesota Data Set

Data from Minnesota were used to examine the methods and results of measuring effect size. These results were compared with estimates from the studies reviewed in the preceding article , "A Practical Method of Policy Analysis by Considering Productivity-Related Research." This section is divided into 13 subsections.

(1) The data set
(2) Simple regression coefficients: the correlation matrix
(3) Partial correlations
(4) Method of analysis: an analytical template
(5) Regression results for teacher/pupil ratio controlled for SES
(6) Comparison with estimates from other studies
(7) Staff qualifications as an intervening variable
(8) Estimating effect size based on "value-added"
(9) Testing the Glass and Smith proposition
(10) Effect size for other staffing categories
(11) Effect size for Minnesota teacher qualifications
(12) Effect size for instructional policy options
(13) Effect size for organizational effectiveness

Table 3
Correlation Matrix

	Math3	*Math5*	*Read3*	*Read5*	*SES*	*Teacher*	*Admin*	*Support*	*Aides*
Math3	1.0000								
Math5	0.7164	1.000							
Read3	0.8693	0.7568	1.0000						
Read5	0.7044	0.9286	0.7929	1.000					
SES	0.6727	0.7574	0.7609	0.8072	1.000				
Teacher	-0.3279	-0.3994	-0.3974	-0.4138	-0.5693	1.000			
Admin	-0.0033	-0.0297	-0.0079	-0.0122	-0.0011	0.0697	1.000		
Support	-0.3256	-0.3245	-0.3288	-0.3394	-0.4025	0.3467	-0.1180	1.0000	
Aides	-0.0312	-0.1197	-0.0708	-0.1030	-0.1307	0.2644	0.1126	0.0148	1.0000

The Data Set

There were some basic problems in estimating effect sizes from the Minnesota data and probably the data from most states. While the achievement scores are by grade level, the number of students and teachers are by school so that individual class sizes cannot be calculated. All other measures are also by school rather than classroom.

The data set in this analysis was constructed for another research project and is described in detail in Phelps (2009). Here I provide a summary. The data set includes 694 elementary schools over a four year period. Achievement is measured for reading and mathematics in the 3rd and 5th grades. There are data related to staffing categories and teacher qualifications. For staffing categories, these include the number of teachers, teacher aides, instructional support personnel, and administrators. Data for teacher qualifications include years of experience, salary, age, and percentage of teachers with Masters degrees. The measure of SES is in the form of an index comprised of five variables as described in Phelps (2009).

Simple Regression Coefficients: The Correlation Matrix

Table 3 presents a correlation matrix produced from the Minnesota data set. The achievement variables are: mathematics scores in 3rd grade (Math3) and 5th grade (Math5); and reading scores in 3rd grade (Read3) and 5th grade (Read5). The data for the staffing categories are measured as the staff/pupil ratio. The observations are:

- Achievement scores are highly correlated by grade and subject.
- SES is highly correlated with achievement.
- All staffing categories are negatively correlated with achievement (higher staff/pupil ratios are associated with lower achievement).
- The staffing categories are positive correlated.
- The high correlation among the staffing category variables (collinearity) poses some complexity in estimating their unique influence on achievement.

Table 4
Partial Correlations

	Math3	*Math5*	*Read3*	*Read5*
Teacher/Pupil Ratio	0.0905	0.0592	0.0671	0.0943
SES	0.5760	0.7016	0.6980	0.7269

Partial Correlations

The partial correlations for the achievement variables tell a different story. When the effect of SES is nullified (partialed out), the correlation between achievement variables and teacher/pupil ratio becomes positive. Table 4 presents the partial correlations, and the "break point," the SES correlation coefficient where the partial correlation of the teacher/pupil ratio is zero. As the SES correlation increases, so does the partial correlation, in this case from a negative sign to a positive sign. Including some measure of SES is critical to any estimate of the influence of class size.

Method of Analysis: An Analytical Template

My original plan was to use a statistical package to run a series of regressions and report the results. This became cumbersome. While there is a great deal of information provided by statistical packages, some is devoted to making probability inferences, and the specific information needed for the policy analysis had to be moved to another setting, in this case a spreadsheet. It was possible to do the statistical calculations for the policy analysis within the spreadsheet itself. A template was created, and only the essential data required for the specific analysis was entered. Consequently, with a correlation matrix, means, and standard deviations for the essential variables, the calculations were processed and presented together in a single spreadsheet format.

Figure 2
Analysis Template to Estimate Effect Size

A	B	C	D	E	F	G	H	I
1		Correlation Matrix			Partial Correlation			
2	Variables	Math5	SES	T/P Ratio				
3	Math5	1.00000			(14.27)			
4	SES	0.75740	1.00000		0.7032			
5	T/P Ratio	-0.39937	-0.56932	1.00000	0.0593			
6								
7	Mean	1484.00	956.00	67.97				
8	Std Dev	87.73	232.00	13.28				
9								
10	Simple Regression: Y = bX + a							
11		Slope (standard partial, Beta)			R Square			
12	(15.11)	SES	0.7574		SES	0.5736		
13		T/P Ratio	-0.3994		T/P Ratio	0.1595		
14		Slope (partial, b)		Intercept (15.7)		Verification: Mean of X must produce Mean of Y)		
15	(15.5)	SES	0.2864	1210.19		1484		
16		T/P Ratio	-2.6383	1663.33		1484		
17	Multiple Regression: Y = b1X1 + b2X2 + a						Common Variance Attributed to:	
18		Slope (standard partial, Beta)		R Square		Calculation	T/P Ratio	SES
19	(16.3)	SES	0.7842	(16.5)	SES	0.59396	0.55634	0.57365
20		T/P Ratio	0.0471	(16.5)	T/P Ratio	0.01881	0.01881	0.00150
21				(16.1)	Total	0.57515	0.57515	0.57515
22		Slope (partial, b)		Intercept (16.4)				
23	(16.2)	SES	0.2965	1179.35		1484		
24		T/P Ratio	0.3111					
25		SEE (Standard error of estimate)						
26	(16.6)	57.1826						

Note: T/P Ratio = Teacher/Pupil Ratio. Std Dev = Standard Deviation.

The analytical template concentrated on the essential calculations for the later policy analysis. The policy model assumed a relationship between the policy option, in this case class size and achievement; therefore, inferential statistics were not critical. What was essential was the estimate of the magnitude of the relationship between achievement and class size, or effect size. Once the template was constructed, it was tested against a standard regression program to assure accuracy. The template consisted of two main parts: (1) Data entry comprised of the correlation coefficients, means, and standard deviations; and (2) calculations producing the regression coefficients, i.e., the weights, or effect sizes.

Statistics were calculated for simple regression (one explanatory variable) and multiple regression, with SES and teacher/pupil ratio as the explanatory variables. Simple regression results begin at B10 on the spread sheet in Figure 2, and multiple regression results begin at B17. Statistics include partial correlation coefficients; standard partial coefficients, or Beta weights; partial coefficients, or "b" weights with intercepts; the R^2, the proportion of explained variance; and standard error of estimate. Several estimates of the R^2 were provided. Verification of the functions is also included. (See G14 on the spreadsheet.) The numbers in parentheses refer to the formulae provided in Appendix A.

Regression Results for Teacher/Pupil Ratio Controlled for SES

The estimated magnitude of the relationships between the four achievement measures (mathematics and reading in the 3rd and 5th grades) and teacher/pupil ratio are presented in Table 5. The effect size estimates are the standard regression coefficients or Beta weights; b-weights with intercept; and R^2, the coefficient of multiple determination. The means of the achievement variables are also provided. From Table 5, the following observations are made:

Table 5
Effect Size Estimates for Teacher/Pupil Ratio

Coefficients	*Read3*	*Read5*	*Math3*	*Math5*	*Mean*
Teacher/Pupil Ratio					
Beta	0.0529	0.0677	0.0815	0.0471	0.0623
R Square	0.0210	0.0280	0.0267	0.0188	0.0236
SES					
Beta	0.7909	0.8457	0.7191	0.7842	0.7850
R Square	0.5597	0.6267	0.4303	0.5940	0.5527
Intercept	1198.25	1176.07	1179.35	1178.62	1183.07
SES	0.2712	0.3425	0.2965	0.2846	0.2987
Teacher/Pupil Ratio	0.3167	0.4789	0.3111	0.5635	0.4176

- SES is by far the most influential variable, explaining over half the variance, 55.27% on average, consistent with many other studies.
- When the teacher/pupil ratio is controlled for SES, the coefficient sign shifts from negative, from the correlation matrix, to positive.
- The higher the correlation between SES and achievement, the larger the teacher/pupil ratio coefficient.
- While positive, the magnitude of the relationship is small, 2.36% of the variance.

Table 6
R^2 Range by Achievement Results: Common Variance Attributed to Teacher/Pupil Ratio

Is common variance attributable to teacher/pupil ratio?	*Read3*	*Read5*	*Math3*	*Math5*	*Mean*
Yes	0.0210	0.0280	0.0267	0.0188	0.0236
No	0.0019	0.0031	0.0045	0.0015	0.0027

Table 7
R^2 Estimates for Teacher/Pupil Ratio and Achievement from Hedges, Laine, and Greenwald (1994)

Beta	0.0176	0.0210	0.0176	0.0114
Estimated r	0.4	0.4	0.4	0.4
Estimated R^2	0.0070	0.0084	0.0070	0.0046

Variance is divided into two parts, the part unique to each variable and the part in common among variables. Therefore, the amount of explained variance depends on whether the common variance is attributed to SES, as is the case in regression,[9] or to teacher/pupil ratio. Table 6 presents the range when the common variance is and is not attributed to teacher/pupil ratio.

The policy implications of these results are clear: Adding teachers has a small effect on achievement. Moreover, the size of the effect depends on the inclusion of an SES variable, the weight of the SES variable, and the attribution of common variance.

Comparison with Estimates from other Studies

Hedges, Laine, and Greenwald provided estimates of the standardized regression coefficients (Betas) for teacher/pupil ratio and four estimates of effect size. These estimates have been converted to R^2 in Table 7 in order to compare them with the Minnesota estimates. The R^2 is calculated from the Beta-weight by multiplying it by the correlation coefficient between achievement and teacher/pupil ratio. The actual correlation is unknown, so a "guess-estimate" of .40 was selected.[10] These estimates are about midway between the high and low estimates from the Minnesota data.

Walberg and the Tennessee STAR experiment (Achilles 1993) provided effect size estimates. These estimates present additional problems because they are effect differences between control and experimental groups rather than standard regression coefficients. Walberg estimated the effect difference at .09 and STAR at about .24. Because there is no measure of the change in the teacher/pupil ratio, a standardized coefficient cannot be calculated directly, but an estimate can be made indirectly. (Beta is a one standard deviation change of achievement for a one standard deviation change in effect.) Assuming a one standard deviation change in the teacher/pupil ratio, the standard regression coefficients (Beta) would be .09 and .24 respectively; assuming a 2 standard deviation change for the STAR project, the Beta would be .12. Assuming a correlation coefficient with achievement of .40, the R^2 is substantially higher than the other estimates.

Table 8
R^2 Estimates from Walberg (1984) and Tennessee STAR Experiment

Studies	*Number of Standard Deviations*	*Difference*	*Correlation Coefficient*	*R Square*
Walberg	1	0.09	0.40	0.036
STAR	1	0.24	0.40	0.096
STAR	2	0.12	0.40	0.048

The Walberg estimate is about double that of the Minnesota estimate and five times higher than the analysis of Hedges et al. The Tennessee STAR estimates are substantially higher than the other two, although the 2 standard deviations assumption puts the estimates in the "ball park." These estimates will be used in the policy analysis to follow.

Staff Qualifications as an Intervening Variable

It might be possible for intervening variables other than SES to have an influence on the estimated magnitude of the class size and student achievement relationship. Data were available to test a teacher qualifications variable. Using the variables average years experience, average salary, average age, and percent of teachers with Masters degrees, a qualifications index was developed to predict mathematics achievement. Regression coefficients were applied to the data from each school to form a single index number representing the influence of these qualifications variables on achievement. The relationship between achievement and teacher/pupil ratio was calculated, including this index, with no change of results; that is, adding a qualifications index to the SES index did not improve the estimate in effect size. Because of the null results, the specifics are not reported here. Once again, the same underlying issue emerged: All variables, including variables related to teacher qualifications, are intercorrelated. Once one of the variables is included in the regression equation, it consumes the common variance and leaves little remaining unique variance for the subsequent variables.

Estimating Effect Size Based on "Value-Added"

Hanushek (2007) advocated a value-added method of production function analysis whereby value-added is achieved by inserting prior years achievement as a lag variable into the regression equation. With regard to the use of a lag variable, he stated: "Clearly, simply estimating relationships between the current level of achievement and the current inputs has little chance of accurately separating the various influences on achievement. Almost certainly, current inputs are correlated with past inputs, leading to obvious problems. The now standard approach on analyzing the growth in student achievement [the lag variable]... substantially reduces the problem" (p.168).

However, there is another consequence. Assuming that the factors influencing achievement are SES, staffing quantity, staffing qualification, and instructional materials (Phelps 2009), these factors

Table 9
Effect Size Estimate for Categories of Staff-to-Pupil Ratios

Staff-to-Pupil Ratios	Math5		
	r	Beta	R^2
Teacher	-0.3994	0.0470	-0.0188
Administrator	-0.3478	-0.0289	0.0009
Support	-0.3245	-0.0234	0.0076
Aide	-0.1197	-0.0211	0.0025
SES	0.7574	0.5940	0.7842

Table 10
Estimated Range of R^2 for Minnesota Teacher Qualifications

Qualifications (expressed as averages)	R^2 Range	
	Low	High
Years of Experience	0.0073	0.0230
Salary	0.0003	0.0007
Age	0.0035	-0.0074
Percent with Masters Degree	0.0000	0.0001

Table 11
Estimated Range of R^2 for Teacher Qualifications from Hedges, Laine, and Greenwald (1994)

Qualifications (expressed as averages)	Beta		Correlation	R^2	
	Low	High		Low	High
Years of Experience	0.0414	0.0550	0.2625	0.0109	0.0144
Salary	0.0366	0.0390	-0.0445	-0.0016	-0.0017
Age	-0.0300	-0.0200	0.1102	-0.0033	-0.0022
Percent with Masters Degree	-0.0300	-0.0200	0.1102	-0.0033	-0.0022

will be present in the lag variable as well as the variables in the last time period. It is easily demonstrated that what is being measured is the difference in factors. Nevertheless, I entered a lag variable into to the regression equations for reading and mathematics at the 5th grade with little additional explanatory power, .0009 for reading and .0147 for mathematics. Because, this value-added method did not add to the measurement of effect size, it was dropped from further consideration in this analysis.

Testing the Glass and Smith Proposition: Does Achievement Improve at an Increasing Rate of Return under a Class Size of 15?

The Minnesota data have schools with class sizes lower than 15, so the Glass and Smith proposition was tested. As class sizes progressed lower than 15, predicted achievement, adjusted for SES, did not increase; in fact, it decreased slightly. It will not be considered further.

Effect Size for Categories of Staff-to-Pupil Ratios

When analyzing categories of staff-to-pupil ratios, such as those for administrators, teacher-support, and teacher-aides, the conclusions are substantially the same as for teachers. The comparison for each of the achievement measures for the four years of data were analyzed in Phelps (2009). Because the results were similar, only the data for one achievement measure, 5th grade mathematics, for one year, is presented here. (See Table 9.) In summary, for staff-to-pupil categories:

- SES explains virtually all the variance.
- The coefficient (Beta) is positive for teachers but negative for all others.
- The additional R^2 for the staffing categories is small, most likely zero for all categories except teachers.

Effect Size for Minnesota Teacher Qualifications

Minnesota data were available for the following categories of teacher qualification: Average years experience; average salary; average age; and average percentage of teachers with Masters degrees. Table 10 presents the R^2 range for these categories.

Using the method described earlier (R^2 = Beta * r), Table 11 presents the estimated R^2 for teacher qualifications from Hedges et al. The Minnesota correlations are used to calculate the R^2 from the Betas. There is a change of sign for salary because of the negative correlation.

Effect Size for Instructional Policy Options

Walberg listed estimated effect sizes for instruction, home influences, and time policies. The effect sizes are actually "effect differences" between a control group and an experimental group, and when added together, they total over 12 standard deviations. Does this mean that if all of the items were implemented by a school at the very bottom of the population (-6 standard deviations), they would progress to the very top (+6 standard deviations)? Surely not! There must be a more practical interpretation. Because of the large number of items, their conceptual similarity, and their likely intercorrelations (shared variance), they are first combined into the categories of curriculum, instructional methodology, instructional

Table 12
Effect Differences and Estimated R^2 for Instructional Categories from Walberg (1984)

	Curriculum	*Method*	*Organization*	*Home*
Average	0.355	0.624	0.113	0.523
Beta	0.118	0.208	0.038	0.174
R^2 (r = 0.5)	0.059	0.104	0.019	0.087
Total R^2	0.269			

organization, and home influences. The average of the effect differences was calculated, reducing the standard deviation range. Second, as a matter of conjecture, two assumptions were made: The treatment difference between the control and experimental group was 3 standard deviations, so the standard regression coefficient (1 Beta) would be one-third the averaged value; and the correlation coefficient with achievement was .5 (R^2 = r * Beta). Based on these assumptions, the revised effect sizes for the categories are listed in Table 12.

With these assumptions, the R^2 are in the range of about .02 to .10, and total to approximately .27. Is there a way to determine if these estimates, or any of the other estimates, are reasonable? The next subsection provides a possible answer.

Effect Size for Organizational Effectiveness

Levin (1997) described the operations of an Accelerated School Program and presented the achievement results.[11] The overall emphasis of the program is on greater organizational effectiveness with the existing resources. For an increase of 1% in expenditures, mathematics achievement increased 45%. The information necessary to calculate an estimated effect size was unavailable although Levin claimed the influence was substantial. He identified two structural elements for consideration in a policy analysis: Incentives linked to successful performance and use of productive technology.

Building on Levin's approach, Phelps (2009) measured the potential effect size attributable to organization effectiveness. From the Minnesota data set, indices were constructed for SES, staff qualifications, staff quantity, and instructional materials. These were entered into the regression equations for the four achievement variables for each of the four years. The residuals were averaged over the four years for each observation to form a new variable, and this variable was entered into the regression equations. This process is a variation of fixed effects estimation in econometrics.[12] Schools consistently either overperformed or underperformed with regard to predicted achievement. The degree by which they missed their target is considered the measure of effectiveness.[13] The analysis also separated district effectiveness from school effectiveness. Because the analysis was of the residual and not actual data, there is no attribution to specific organizational behaviors. See Table 13 for the effect size estimates.

These estimates are valuable for several reasons:

- The measure of effectiveness--averaging of the residuals over time--substantially reduces the error variance of the equations to 0.075.
- The estimates provide an empirical base for the boundaries of effect size for the various categories of policy options described above. First, the resource-oriented variables such as staffing quantity (class size), staff qualifications (built into the salary schedules), and instructional materials seem to be limited in their overall contribution to around the average of .063. Second, the instructional and organizational variables as suggested by Walberg and Levin, do not appear to exceed the effectiveness total of .285. (The "guess-estimate" made earlier was .269.)
- The data suggest differences in the contribution of the resources and effectiveness variables based on subject matter; resources could be more important for reading, while effectiveness more important for mathematics.
- Effectiveness appears to be a shared responsibility between school and district policies and operations. This seems to imply that skilled district staff might be helpful in providing individual schools with instructional and management assistance. Moreover, good district policies would seem to support good policies in schools.

Summary and Conclusion

In this article, several achievement production models were identified stressing the importance of a unified and comprehensive operating structure, and quantifiable relationships among the elements of the structure. The studies reviewed here do not typify either a comprehensive structure or consistent measure of effect size. Based on the previous evidence and arguments presented, a fresh model emerges which provides a unifying structure, a consistent method of estimating effect size, and a coherent set of assumptions. This model emphasizes an effect ceiling and organizational effectiveness.

Table 13
Effect Size for School and District Effectiveness

Student Achievement	*Without Residual*	*SES*	*Indices*	*District Effectiveness*	*School Effectiveness*	*Total*	*Error*
Mathematics	0.585	0.550	0.035	0.185	0.155	0.340	0.075
Readings	0.710	0.620	0.090	0.120	0.110	0.230	0.060
Mean	0.648	0.585	0.063	0.153	0.133	0.285	0.068

The effect ceiling requires a different way of measuring effect size, while the inclusion of effectiveness variables substantially increases the accuracy of prediction. Most importantly, the model brings a new policy focus to the dilemma of Hedges, Laine, Greenwald: Why focus the primary attention on merely increasing resources (expenditures or reducing class size) if substantial achievement benefits can be derived from better instructional and organizational policies?

A Unified Structure

The reviewed research in this article focused mostly on small components of the educational process rather than treating the components as elements of a comprehensive unified structure. Class size is the primary center of attention while staffing categories other than teachers are largely ignored, counter to the notion of a team of people working together. The individual components of teacher qualifications also are viewed separately, instead of working together. Individual components of the instructional program, such as curriculum, methods, time, and instructional materials, are also viewed separately. In every case, the components are not unique or isolated; instead they are conceptually, operationally, and statistically related. An enhanced understanding of educational organizations comes from a paradigm encompassing a comprehensive system rather then reductionism to individual components.

Viewing education as a comprehensive system has implications for policy analysis. By identifying the larger categories of education and having estimates of their contribution, as well as the contribution of the component elements, it is possible to model the operation of the entire system. By simulating changes in multiple policies, the model estimates change in multiple achievement outcomes.

A unified educational structure, with its quantifiable component elements, is described in Table 14. This paradigm allows for expansion and modification of the structure to fit any circumstance where effect size and incremental cost of the policy options can be estimated. The structure that will be used in the simulation model described in the next article, "A Practical Method of Policy Analysis by Simulating Policy Options," is:

Achievement = SES+Staff Quantity+Staff Qualifications+
Instructional Program+Organizational Effectiveness

Estimating School-Specific Effect Size

The major consequence associated with the variance measure of effect size is its school-specific nature. Because the variance measure of effect size is a curve, every school will have a unique position on the curve; that is, every school will have a different marginal effect size depending on its unique circumstance. Estimating the potential of the policy options is based on seven major principles. Each principle has a different role in determining the most cost-effective policy options for the school.

Principle 1: Role of effect size. Good policy decisions start with good strategies. What is to be accomplished? How is it to be accomplished? Who is responsible? What training and mentoring is required? How will the performance and progress be monitored? Reducing class size or adding staff without first addressing these questions is foolhardy. In essence, merely adding staff without clear and comprehensive instructional (Walberg) and organizational (Levin, Phelps) strategies is counterproductive.

Principle 2: Accommodating uncertain effect size. The measurement of effect size is not precise, and research provides little in the way of reliable measures.[14] However, not all is lost. Ranges of effect sizes can be used to separate weak policy options from those with stronger possibilities. If there is a good strategy in place, then it is reasonable to assume the maximum effect size could be realized. Without a strategy, the minimum effect size is a more reasonable assumption.

Principle 3: Role of distribution variance. If effect sizes of two policy options are virtually equal, the policy with the largest variance will have the greater potential. The ability to predict is proportional to the variance; variables with larger variance are better predictors than variables with smaller variance. Other things being equal, weight should be given to the policy with the larger variance.

Principle 4: Role of the school's current status. An underlying assumption of this conceptualization is the notion of a ceiling effect—after a point, benefits for the policy option diminish. The "benefit curve" is an S-shaped curve with achievement on the Y-axis and the policy variable on the X-axis. If a school's position is low on the policy variable, the potential for improved achievement gradually increases. In contrast, if the school's position is high on the policy variable, the potential for improvement gradually diminishes.

Principle 5: Nonincremental policy options. Some policies are binary, not distributional. For example, if a new mathematics or science curriculum is based on a textbook, the policy is binary—either the textbook is adopted or it is not. Therefore principle 4 does not apply and a different method is required, which will be discussed in the next article.

Table 14
Quantifiable Component Elements of a Unified Educational Structure

Student Achievement	*SES*	*Staff Quantity*	*Staff*	*Instruction*	*Effectiveness*
Early Grades Reading Mathematics	Unique to each state	Teachers Support Aides Administration	Qualifications Education Experience Salary	Curriculum Methodology Organization Homework Time Technology	School District

Principle 6: Estimating the marginal cost-effectiveness. There are three necessary numbers required to calculate the marginal cost-effectiveness of any policy option: the estimated effect size; the incremental cost; and the Z-score on the policy variable.[15] The calculation is: Effect-Size times School-Position times Marginal-Cost times.

Principle 7: Role of cost-effectiveness. If the effect sizes of two options are virtually equal, the policy with the least cost is the most cost-effective. In a complicated situation such as schools, these hand-calculations would be virtually impossible. However with current computer software, these calculations are made within fractions of a second.

References

Achilles, C.M., B.A. Nye, J.B. Zaharias, and B.D. Fulton. "The Lasting Benefits Study (LBS) in Grades 4 and 5 (1990–1991): A Legacy from Tennessee's Four-year (K–3) Class-size Study (1985–1989)." Project STAR. Paper presented at the North Carolina Association for Research in Education, Greensboro, North Carolina, January 14, 1993.

Addonizio, Michael F., and James L. Phelps. "Class Size and Student Performance: A Framework for Policy Analysis." *Journal of Education Finance* 26 (Fall 2000): 135-156.

Bohrenstedt, George W., and, Brian M. Stecher, eds. *Class Size Reduction in California: Early Evaluation Findings, 1996-98*. CSR Research Consortium. Sacramento, CA: California Department of Education, June 1999. http://www.classize.org/techreport/index.htm.

_____. *Capstone Report: What We Have Learned about Class Size Reduction in California*. CSR Research Consortium. Sacramento, CA: California Department of Education, August 2002. http://www.classize.org/techreport/index-02.htm.

Forrester, Jay W. "Systems Dynamics: Future Opportunities." In *Studies in Management Science*, edited by Augusto A. Legasto, Jay W. Forrester, and James M. Lyneis, 7-21. Vol. 14. Amsterdam, North Holland: System Dynamics, 1980.

Glass, Gene V., and Mary Lee Smith. *Meta-analysis of Research on the Relationship of Class-size and Achievement*. San Francisco, CA: Far West Laboratory for Educational Research and Development, 1978.

Guilford, Joy Paul. *Fundamental Statistics in Psychology and Education*, 4th ed. New York: McGraw-Hill, 1965.

Hanushek, Eric A. "Some U.S. Evidence on How the Distribution of Educational Outcomes Can Be Changed." In *Schools and the Equal Opportunity Problem*, edited by Ludger Woessmann and Paul Peterson, 159-190. Cambridge, MA: MIT Press, 2007.

Hedges, Larry V., Richard D. Laine, and Rob Greenwald. "Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes." *Educational Researcher* 23 (April 1994): 5-14.

Krueger, Alan B. "Understanding the Magnitude and Effect of Class Size on Student Achievement." In *The Class Size Debate*, edited by Lawrence Mishel and Richard Rothstein, 7-35. Washington DC: Economic Policy Institute, 2002. http://edpro.stanford.edu/hanushek/admin/pages/files/uploads/classsizedebate.full%20volume.pdf.

Kuhn, Thomas S. *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press, 1970.

Levin, Henry M. "Raising School Productivity: An X-Efficiency Approach." *Economics of Education Review* 16 (June 1997): 303-311.

Mosteller, Frederick. "The Tennessee Study of Class Size in the Early Grades." *Future of Children* 5 (Summer/Fall 1995): 113-127. https://www.princeton.edu/futureofchildren/publications/docs/05_02_08.pdf.

Phelps, James L., "Optimizing Educational Resources: A Paradigm for the Pursuit of Educational Productivity," *Educational Considerations* 35 (Spring 2008): 3-18.

_____. "Measuring and Reporting School and District Effectiveness." *Educational Considerations* 36 (Spring 2009): 40-52.

Schrage, Linus E. *Lindo: An Optimization Modeling System*. San Francisco: CA: Scientific Press, 1991.

Walberg, Herbert J. "Improving the Productivity of America's Schools." *Educational Leadership* 41 (May 1984): 19-27.

Williams, H. Paul. *Model Building in Mathematical Programming*, 2nd ed. New York: Wiley and Sons, 1985.

Wooldridge, Jeffrey M. *Introductory Econometrics: A Modern Approach*. Cincinnati, OH: South-Western College Publishing, 2000.

Endnotes

[1] Schools have no control over the number of students, only the number of teachers.

[2] All subsequent references to Walberg in this article refer to Herbert J. Walberg, "Improving the Productivity of America's Schools," *Educational Leadership* 41 (May 1984): 19-27.

[3] The lack of a meaningful measure of SES may explain why the results from studies regarding teacher/pupil ratios and achievement are so diverse.

[4] Glass and Smith (1978) assumed an increasing return to scale and used a squared term to achieve that result. The model produced a curve with an increasing and decreasing return to scale, so they made an adjustment transforming the decreasing return to a consistent return to scale.

[5] See Phelps (2008). See also, section 3, Appendix A of this article.

[6] See the comments in the preceding article, "A Practical Method of Policy Analysis by Considering Productivity-Related Research," and Phelps (2009). This is called a fixed effect in econometrics. See also, Wooldridge (2000).

[7] All subsequent references to Hedges et al. in this article refer to Larry V. Hedges, Richard D. Laine, and Rob Greenwald, "Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes," *Educational Researcher* 23 (April 1994): 5-14.

[8] All subsequent references to Glass and Smith in this article refer to Gene V. Glass and Mary Lee Smith, *Meta-analysis of Research on the Relationship of Class-size and Achievement* (San Francisco, CA: Far West Laboratory for Educational Research and Development, 1978).

[9] The variable with the highest correlation consumes the common variance.

[10] A correlation of .4 is similar to the Minnesota data, although the sign was negative in the Minnesota case.

[11] All subsequent references to Levin in this article refer to Henry M. Levin, "Raising School Productivity: An X-Efficiency Approach," *Economics of Education Review* 16 (June 1997): 303-311.

[12] See Wooldridge (2000).

[13] It is analogous to rolling a die: Some schools consistently rolled 1, 2, and 3, while others rolled 4, 5, and 6, with the target of 3.5, the average.

[14] According to Schrage (1991, 8), "The first rule of modeling is don't waste time accurately estimating a parameter if a modest error in the parameter has little effect on the recommended decision."

[15] The Z-score determines where the school is positioned on the S-shaped curve.

[16] The source for these formulae is Joy Paul Guilford, *Fundamental Statistics in Psychology and Education*, 4th ed. (New York: McGraw-Hill, 1965). Related page numbers are in parentheses.

[17] Note that the value of the correlation coefficient with the same subscript numbers, e.g., r22, is 1.

Appendix A

1. Formulae for estimating effect size

Following are the formulae used to calculate the statistics in the template.[16]

1.1 Partial Correlation (14.27, p. 339):

$$r_{12} = r_{12} - r_{13}\, r_{23} / \sqrt{(1- r^2_{12})(1-r^2_{23})}$$

1.2 Coefficient for linear regression (15.55, p. 367):

$$b_{yx} = r_{yx}\, (\sigma_y / \sigma_x)$$

1.3 The "a" coefficient in a linear regression equation (15.7, p. 368):

$$a = M_y - (M_x)b_{yx}$$

1.4 Relation of regression coefficients to r2 (15.9, p. 368):

$$b_{yx}\, b_{xy} = r^2$$

1.5 Regression equation with standard measures (15.11, p. 370):

$$Z_y = r_{yx}\, Z_x$$

1.6 Standard error of estimate (15.16, p. 373):

$$\sigma_{yx} = \sigma_y \sqrt{(1- r^2)}$$

1.7 Square of coefficient of multiple correlation with three variables: (16.1, p. 394).

$$R^2 = r^2_{12} + r^2_{13} - 2r_{12}\, r_{13}\, r_{23} \;/\; 1 - r^2_{23}$$

1.8 Partial regression coefficients, the "b" weight (16.2, p. 396):

$$b = (\sigma_1/\sigma_2)\;\; \beta_{12}$$

1.9 Standard partial regression coefficients (16.3, p. 396):

$$\beta_{12} = r_{12} - r_{13}\, r_{23} / 1 - r^2{}_2 3$$

1.10 The "a" coefficient for linear regression (16.4, p. 397):

$$a = M1 - b_{12}M_2 - b_{13}M_3$$

1.11 Calculating the multiple R from Beta coefficients (16.5, p. 39):

$$R^2 = \beta_{12}r_{12} + \beta_{13}r_{13}$$

Note that if the correlation is negative, the absolute value is taken. However, the result is not consistent with equation 16.1. Actually the R^2—the proportion of explained variance—is divided into two parts, the unique part and a common part. Equation 16.5 attributes both the unique and common parts to each variable, thus the sum is larger than 16.1. As a result, a choice must be made as to which variable will receive the common variance. The unique variance of the remaining variable is calculated by subtracting the unique and common variance of the selected variable from the R^2 from equation 16.1:

$$R^2 - \beta_{12}R_{12} = \beta_{13}R_{13}$$

This is consistent with the principles of stepwise regression. The first term in (with the highest correlation with the outcome variable) assume both the unique and common variance with the other variables. The next variable in assumes just the unique variance.

1.12 Standard error of multiple estimate (16.6, p. 400):

$$\sigma_{yx} = \sigma_y \sqrt{(1- R^2)}$$

1.13 Multiple regression with more than three variables (16.13, p. 409)

Each time a variable is added to the regression equation, the Betas must be recalculated. The calculation answers the question: What regression weights would best predict the outcome variable from the explanatory variables? The calculation is based on normal equations, with one fewer equation than the number of variables in the equation (including the outcome). The solution to these normal equations can be found by employing a software program, like Microsoft Excel's Solver. The follow example can be expanded to include any number of variables.

$$r_{22}\, \beta_{12} + r_{32}\, \beta_{13} + r_{42}\, \beta_{14} = r_{12}$$
$$r_{23}\, \beta_{12} + r_{33}\, \beta_{13} + r_{43}\, \beta_{14} = r_{13}$$
$$r_{24}\, \beta_{12} + r_{34}\, \beta_{13} + r_{44}\, \beta_{14} = r_{14}$$

2. Converting standard regression coefficients to R^2

The following principles apply. If a value is unknown, then an estimate must be made to stay within the principles.

2.1. The total of all the variance is 1: $R^2 = 1$

2.2. The R^2 for the individual explanatory variables is calculated by the formula:

$$R^2 = \beta_{12}r_{12} + \beta_{13}r_{13} + \ldots\, \beta_n r_n$$

2.3. The estimated range of the nonresource explanatory variables is:

SES = 55 to 60; Error 7 to 10

Effectiveness (instructional and organizational) 25 to 27.

2.4. The range for the resources explanatory variables, therefore, must be between 3 and 13.

3. Interpretation of Variance

Statistical variance is a general term referring to the area under the normal distribution, but it is measured in two ways. The first method is in terms of square units, and the second is in terms of a linear parameter of the normal distribution. It is important to distinguish between the two measures because the same word, variance, is used to describe both concepts. The focus here is on how variance can be the bases of estimating effect size.

3.1. The sum of squared deviations from the mean of the distribution gives a measure of the total area under the distribution, or total variance area.

3.2. The parameter of the distribution is calculated by taking the average squared deviation, also called the variance, or σ^2, the square root of which is the standard deviation or σ. The standard deviation is the width parameter of the distribution. The standard deviation is also the parameter in determining the area under the normal curve: $\sigma\sqrt{2\pi}$.

3.3. The principle of regression is to find a line for which the sum of the squared deviations (area) around the line is a minimum. This is the error variance area. Because the regression line is the mean of the distribution, the standard error of estimate is the standard deviation or width parameter of the distribution around the line (p.375). In other words, the total variance area is comprised of the explanatory variance area and error variance area.

3.4. Divided equation (3.3) by the total variance area, the results are percentages, the percentage attributable to the explanatory variables and error. Because the total percentage is 1.00, the percentage of the explanatory variance area (that explained by the regression line) and error variance area are:

1 = % Explanatory Variance Area + % Error Variance Area

3.5. Regression programs provide these sum of the square numbers from which the explanatory variance area is calculated. It is said the explanatory variable explains a certain proportion of the total variance. It is called the coefficient of determination, and noted as the R^2.

3.6. Each explanatory variable has a unique R^2 based on the relationship between the Beta and correlation coefficient:

$R^2 = \beta_{12}r_{12} + \beta_{13}r_{13}$

3.7. As additional explanatory variables are added, as is the case in stepwise regression, the amount of explanatory variance increases to a maximum point.

3.8. The area of the normalized curve is 1; therefore the proportion of variance explained by each component, explanatory variables and error (or residual), sum to 1.00 with the R^2 for each component representing a percentage of area under the normal curve.

3.9. The percentage area of each component can be converted to the cumulative area under the normal curve or percentile. This curve is S-shaped with asymptotes at 0 and 100 percentiles. Because the mean of the explanatory variable equals the mean of the outcome variable, one-half of the R^2 area is above the 50th percentile and one-half below. For example, if the R^2 is .50, the asymptotes are at the 25th and 75th percentiles.

4. Calculations for the normal curve and area under the curve

4.1 The equation for the normal curve is:

$Y = e \exp -Z^2/2 / \sqrt{} (2 Pi)$

The cumulative area under the normal curve is the integral of the normal curve. Therefore, the slope of the integral at any point is calculated via the normal curve equation by inserting the value of Z.

Appendix B

Summary of Effect Sizes Converted to R^2

Tables B-1 and B-2 summarize the materials presented in the body of this article. In Table B-1, the effect sizes are presented in terms of the amount of variance explained or the R^2. In some cases, a conversion was made from the original metric to the R^2 metric, based on the formulae described previously. The summary is presented in three major categories: Staffing; instruction; and qualifications. Each of the categories includes the associated elements. For each of the studies reported, a low and high estimate are presented. When the correlation or Beta coefficient is negative, the results are presented as negative.

In Table B-2, summary calculations are provided. For each category and element an average low, average high, and average are calculated. In order to evaluate the estimates, the absolute values are calculated and then totaled to determine their total explanatory value, the total of which cannot exceed 1.00, including error. The Staffing category ranged from .0437 to .0587; Instruction ranged form .1523 to .2700; and Qualifications from .0178 to .0240. The totals for these categories ranged from a low of .1870 to a high of .3527, with the average of .2640. When the R^2 of SES is set as .5800 (from the Minnesota data), the error contribution is calculated.

When these data are taken together, the ranges are similar to the results obtained from the analysis of the Minnesota data set. Importantly, these data reflect the product of a methodology to estimate a consistent effect size from studies with different measures. These are not intended to represent a definitive estimate. Nevertheless, these estimates are thought to be a reasonable starting point for use in a simulation model.

Summary Table B1
Effect Sizes from Various Studies

Variables	*Minnesota*		*Hedges et al.*		*Krueger*		*Walberg*		*STAR*		*California CSR*	
	Low	*High*	*Low*	*High*	*Low*	*High*	*Low*	*High*	*Low*	*High*	*Low*	*High*
Staffing												
Teacher/Pupil Ratio	0.0015	0.0188	0.0070	0.0080	0.0800	0.0800	0.0360	0.0450	0.0400	0.1000	0.0000	0.0400
Support/Pupil Ratio[1]	-0.0076	0.0005										
Aide/Pupil Ratio	-0.0025	0.0004							0.0000	0.0000		
Administrator/Pupil Ratio	-0.0032	0.0001										
Instruction												
Curriculum							0.0235	0.0470				
Method							0.0415	0.0830				
Organization							0.0015	0.0030				
Homework							0.0350	0.0700				
Time							0.0255	0.0510				
Qualifications												
Experience	0.0073	0.0230	0.0109	0.0144								
Salary	0.0003	0.0007	-0.0016	-0.0017								
Masters Degree	0.0000	0.0001	-0.0033	-0.0022								
Age	0.0035	-0.0074										

[1] "Support" refers to instructional support personnel such as reading teachers.

Appendix B (continued)

Table B2
Summary Calculations

Variables	Average		Average	Absolute Value	Subtotal	Absolute Value		Subtotal
	Low	High				Low	High	
Staffing								
Teacher/Pupil Ratio	0.0329	0.0584	0.0380	0.0380		0.0329	0.0584	
Support/Pupil Ratio[2]	-0.0015	0.0001	-0.0036	0.0036		0.0015	0.0001	
Aide/Pupil Ratio	-0.0013	0.0002	-0.0005	0.0005		0.0013	0.0002	
Administrator/Pupil Ratio	-0.0032	0.0001	-0.0016	0.0016		0.0032	0.0001	
					0.0437			0.0587
Instruction								
Curriculum	0.0295	0.0590	0.0353	0.0353		0.0295	0.0590	
Method	0.0520	0.1040	0.0623	0.0623		0.0520	0.1040	
Organization	0.0100	0.0200	0.0023	0.0023		0.0100	0.0200	
Homework	0.0435	0.0870	0.0525	0.0525		0.0435	0.0870	
Time			0.0383	0.0383				
Qualifications								
Experience	0.0091	0.0187	0.0139	0.0139		0.0091	0.0187	
Salary	-0.0007	-0.0005	-0.0006	0.0006		0.0007	0.0005	
Masters Degree	-0.0017	-0.0010	-0.0013	0.0013		0.0017	0.0010	
Age	0.0017	-0.0037	-0.0020	0.0020		0.0017	0.0037	
					0.0178			0.0240
		Subtotal		0.2520	0.2138	0.1870	0.3527	0.3527
		SES			0.5800	0.5800	0.5800	
		Total			0.7938	0.7670	0.9327	
		Error			0.2062	0.2330	0.0673	
		Grand Total			1.0000	1.0000	1.0000	

[2] "Support" refers to instructional support personnel such as reading teachers.

A Practical Method of Policy Analysis by Simulating Policy Options

James L. Phelps

This article focuses on a method of policy analysis that has evolved from the previous articles in this issue.[1] The first section, "Toward a Theory of Educational Production," identifies concepts from science and achievement production to be incorporated into this policy analysis method. Building on Kuhn's (1970) discussion regarding paradigms, the second section, "Characteristics of an Achievement Production Theory and Model," describes a comprehensive, coherent, and unified theory and a mathematical model of achievement production substantially different from other theories and models. Using sample data, section three, "Example of the Policy Analysis Model," demonstrates the implementation of the model.

Toward a Theory of Educational Production

An Example of Scientific Method

To follow is a brief history of the scientific theory of gravity drawn from Feynman (1965, 17-20). In many ways, it parallels the motivation for and execution of the articles in this special issue. In addition, it highlights some fundamental differences in theory and models between the physical sciences and achievement production.

In ancient times, people believed that the planets circled the earth because earth "just had to be" the center of the universe. Later, Copernicus observed the planets moving in the sky and thought the planets, including earth, moved around the sun. The follow-up questions were: What pattern of motion do the planets follow—a circle or some other curve; and how fast do they move? Tycho Brahe thought he could help answer these questions by carefully recording how the planets move in the sky. From these data, alternative theories explaining the movement were developed. In essence, science was in transition from a philosophy to the collection and analysis of observations in order to develop better explanations.

James L. Phelps holds a Ph.D. from the University of Michigan in Educational Administration. He served as Special Assistant to Governor William Milliken of Michigan and Deputy Superintendent in the Michigan Department of Education. Active in the American Education Finance Association, he served on the Board of Directors and as President. Since retirement, he spends a great deal of time devoted to music, composing and arranging, playing string bass in orchestras and chamber groups, as well as singing in two choirs. He resides with his wife, Julie, in East Lansing, Michigan.

Kepler analyzed the observations made by Brahe and developed three propositions: The planet orbits are in the form of an ellipse; equal areas are swept in equal times; and the time it takes to go around the sun is based on a well-defined mathematical function. Meanwhile, Galileo, while testing the laws of inertia (rolling balls down an inclined plane), concluded that objects always move in a straight line unless some other force acts upon them. The force acting on the planets, Newton concluded, was gravity. The relationship is defined by his mathematical function: $F = G\ m_1m_2/r^2$.

As the ability to make accurate measurements increased, the tests of Newton's theory of gravity became more stringent. Indeed, the movement of the planets and moons could be accurately predicted by his mathematical function. Once the Newton law of gravity was confirmed through experiment, it was possible to build upon that knowledge to develop new knowledge. Based on the same mathematical function, Cavendish was able to determine the value of G, or "weighing the earth," through a laboratory experiment. Einstein later modified the Newton formulation when he discovered that energy and mass were related ($E = MC^2$); light would react to gravity and there is a "cosmic speed limit," the speed of light. The theory of gravity is tested every time an object is sent into space because the values within the equation change—there is a different set of initial conditions.

Still the theory of gravity is not complete. Physicists know that the laws on a small scale (the atomic level) are much different than the laws on a large scale (the universe). The analogy that the electron orbits the nucleus of the atom as the planets orbit the sun is incorrect. The Newton laws as modified by Einstein can predict with great accuracy the position and motion of the planets today and well into the future. On the other hand, there is no law predicting the position and motion of an electron in an atom. Quantum mechanics is built on what is called the "uncertainty principle"; that is, the position and motion of a particle cannot be accurately measured at the same time, but the probabilities can be measured with great accuracy. Today's sophisticated electronics are based on knowing these probabilities. A particle has even been named that controls all the movement in the universe—the Graviton—but to-date no one has been able to detect the particle and measure its properties. The endeavor to develop a complete theory of gravity is likely to be an endless journey.

There are several relevant points from the evolution of gravity theory:

- Over a long period of time, the thinking gradually shifted away from philosophy and beliefs to a science of observation, theory, and experiment. Once a theory was developed from observations, it was tested and verified by experiment. When the experiments more accurately predicted results, the old theories were replaced.
- A basic law can be expanded from the very simple situation to the very complex, e.g., the path of a thrown baseball to the motion of all the objects in the universe.
- The basic law demonstrates that all variables are not of equal influence. It is not necessary for every aspect of the complex system to be considered, only the most important. For example, an object with a small mass and a great distance from the earth (r^2) has virtually no influence of the orbit.

- With the basic law in hand, estimates of other variables within the system are possible. For example, Cavendish measured the coefficient of gravity, G in the formula, by suspending two balls from strings and measuring their attraction.
- With a strong theory behind the basic law, the theory gives direction to future research. In this way, the theories become more sophisticated over time.

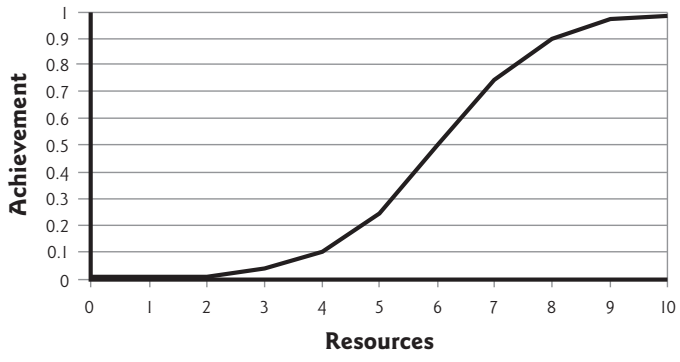
The theory of gravity makes an interesting prediction. If the sun were to suddenly explode, reducing the mass, what would happen to the orbit of earth? Clearly the force would change, and the earth's orbit would change. There would also be other severe consequences. While the change of force would be automatic, the change would not be instantaneous. Rather, it would take about eight minutes—the speed of light—before earth would respond. By some magical and unknown process, "mother nature" knows exactly what to do. How does this scientific example apply to achievement production?

Shortcomings of Current Achievement Production Theory and Modeling

As seen from the gravity example, theories and mathematical models are representations of a phenomenon. Therefore, theories and models must be judged based on how well they characterize the phenomenon and how well they predict events, not based on a how well they reflect people's beliefs. Based on these criteria, there are some apparent shortcomings in the current achievement production theory and models.[2]

While each piece of class size research referenced in earlier articles in this issue has a research question, there is no fundamental theory being tested. What is implied is a "common-wisdom" theory: Reduced class size will *automatically* cause teachers to provide students with greater individual attention and, as a result, achievement will increase. This is not a testable theory. In order for a theory to be tested, it must be sufficiently concrete to allow observational data to be collected and analyzed. The individual attention theory is ill-defined, raising ambiguity regarding the actual theory being tested in class size research. What is implied by individual attention is a theory of changed behavior: By changing the class size or adding any type of instructional staff, staff behavior will automatically change, and so will the behavior of the students. As a result of these changes in behavior, achievement will improve. Before achievement can be expected to change, two critical steps must be taken; and neither step is included in the current theory or mathematical model. First, there must be a change in behavior by the instructional staff, and, second, there must be a change in the behavior of the pupils. The "automatic-individual-attention" theory and interpretation of the current achievement production model is not an accurate representation of the achievement-producing process. More likely, the theory involves a sequence such as a change of policies, a change of teacher and student behaviors, the practice of the new behaviors over time, and only then, a change in achievement.

Figure 1
The Learning Curve



There is another apparent shortcoming of current theory and modeling. According to learning theory and research, achievement does not change at a constant rate especially when there is an upper performance limit, i.e., a perfect score. There is a mathematical model representing the theory developed from observation and analysis: Achievement growth is proportional to the existing achievement level and to the difference between the existing level and the upper limit. (See Appendix B.) This model is in the form of a learning curve, illustrated in Figure 1. By assuming a constant rate of change, most achievement production research does not take the learning theory or the growth model into consideration. Indeed, there is no learning theory supporting a linear relationship between achievement and policy variables; there is only a statistical model with a linear feature. Most productivity research with the relationships proposed by Glass and Smith (1978),[3] i.e., increasing return to scale, and Hedges, Laine, and Greenwald (1994),[4] i.e., a constant return to scale, are inconsistent with this learning curve, and not an accurate representation of achievement growth.

Current achievement production research is mostly designed to test the hypothesis: Do resources (money or class size) make a difference? Studies are generally designed with one explanatory variable (expenditures or class size) and other control variables (e.g., socioeconomic status) and a statistical model to produce a kindly result. If the results are statistically significant, the policy implication is to "invest." In the cases of Glass and Hedges et al., they openly conclude that resources make a difference, and more resources make more of a difference.[5] Over a period of time, and partly due to these studies, a belief system was enhanced. Following this belief system, states and schools districts proceeded to make large investments in lowering class size.

Finally, current theories and models do not provide for the effective implementation of organizational or instructional policies. Because behavior does not change automatically, schools must rely on thoughtful policies as instruments of behavioral change. Since data are not collected regarding such policies, and little is known about their characteristics, these features are usually omitted from research efforts. There is evidence that organization behavior is consistently associated with academic performance and accounting for this behavior substantially increases the ability to predict achievement (Phelps 2009). Therefore, class size, organizational and instructional policies, and effective implementation of the policies all contribute to academic achievement. Theories and models not addressing the role of policies and behavior, the learning curve, or effectiveness do not fully characterize the complexity of achievement production. As a result, the models are less accurate in predicting achievement.

Characteristics of an Achievement Production Theory and Model

This section describes an achievement production theory and model with characteristics evolving from what are considered shortcomings of existing achievement theories and models. It also describes the steps for its implementation. Most importantly, achievement is a complex and dynamic system, which does not behave according to the physical laws determined by "mother nature." Just as a "gravity law" passed by Congress will not automatically change the behavior of the objects in the universe, the mere allocation of resources will not automatically result in improved achievement. While legislatures can allocate funds, they cannot change the shape of the "learning curve" or guarantee the effective use of the funds. In short, the achievement production model must be consistent with how schools teach and how students learn. It also must take into consideration the effective use of resources. This section is divided into three subsections: A policy-behavior-achievement (PBA) theory; the PBA model; and the PBA production model process, with steps for implementation.

A Policy-Behavior-Achievement (PBA) Theory

Because policy is the primary instrument influencing organizational behavior and behavior influences achievement, the proposed theory is: *Educational achievement is the product of all policies influencing staff, community, and student behavior and the effective implementation of those policies*.

There are several categories of policy variables, each with unique characteristics. Each of the categories influences some aspect of behavior.

- Resource or purchased variables include staffing quantity, staffing qualifications, instructional materials, and possibly special facilities.
- Family and community variables are represented by socioeconomic status (SES), which is divided into: the proxies used for measuring the association with achievement, but are beyond the control of schools, e.g., number of students receiving free and reduced-price meals, family income, and parent education; and the usually unmeasured behaviors which are also associated with achievement but are partially under the control of schools and community, such as motivation, discipline, and leisure reading.
- Process or effectiveness variables are organizational, per Levin (1997)[6], and instructional, per Walberg (1984).
- Incentive policy variables include extrinsic and intrinsic rewards for performance.[7]

The important role of behavior in achievement productivity is self-evident when looking at achievement at different organizational levels. Between school districts, there could well be differences in funding and class size accounting for the differences in achievement. Between school buildings within the same school district, the difference in funding and class size would most likely be less; thus the influence on achievement would be less. At the classroom level, there is no difference in funding or class size, but the achievement differences among students is still substantial. The different behaviors of the teacher, student, and family undoubtedly contribute to these achievement differences. This point is missing from other theories and models of achievement production. The contribution of behavior in response to policies is a key component of the policy-behavior-achievement paradigm.

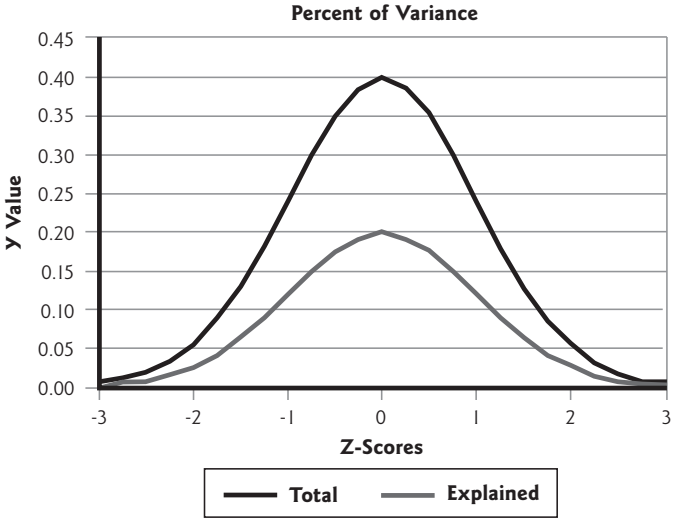
The family and community variable, SES, deserves special attention because of its potential role in influencing behavior. There is no fixed definition of SES. It is a concept for which proxy data are substituted, e.g., percent of students receiving free or reduced-price meals as a proxy for family income. Other proxies are common as well, e.g., parent or community education levels, student mobility, and attendance. In reality, these variables have no direct relationship with achievement. Instead, they are proxies for unobserved behaviors associated with achievement such as parent encouragement, time devoted to reading or homework, and rewards to do well in school. While the school cannot hope to change these proxy variables, it is possible through policy actions to influence the personal behaviors thought to be associated with achievement. This behavior aspect of the family and community variables is accommodated within the model.

It is possible to direct policies toward the educational staff, students, families, and in some cases, the community. In this context, a policy means a course of action to provide direction, assistance, supervision, evaluation, and rewards. An inventory of the various policies across the three groups of recipients will most likely reveal a disproportionate attention to what students should do. Less attention is paid to the instructional staff and little to families and the community, even though the benefits from such polices could be substantial. Because of attitudes regarding academic freedom to teach, or a reluctance to become involved in community and family affairs, a substantial potential may be missed.

Below is a succinct statement of the PBA theory:

- Achievement is the product of many behaviors: The student to study; the school staff to teach; and the family and community to provide a supporting environment.
- Behaviors are influenced by policies: What content the student studies and how they study; what content the school teaches and how the content is taught; and what contribution the family and community make to the educational process. (Learning does take place outside of the school setting.)
- The policies work in combination: Many complementary behaviors are required to produce or improve achievement.
- Some policies are more effective than others, and schools implementing more effective policies produce better academic performance.
- Effective policies can be different for various academic subjects and grade levels.
- Implementing some policies is more cost-efficient than others.
- In order to improve achievement, ineffective policies must be changed, and effective policy must be enhanced.
- Even effective policies eventually reach a point of diminishing returns.
- It is the responsibility of policymakers—school leadership, instructional staff, families, community—to select and implement the most cost-effective policies.

Figure 2
Total and Explained Variance in Student Achievement



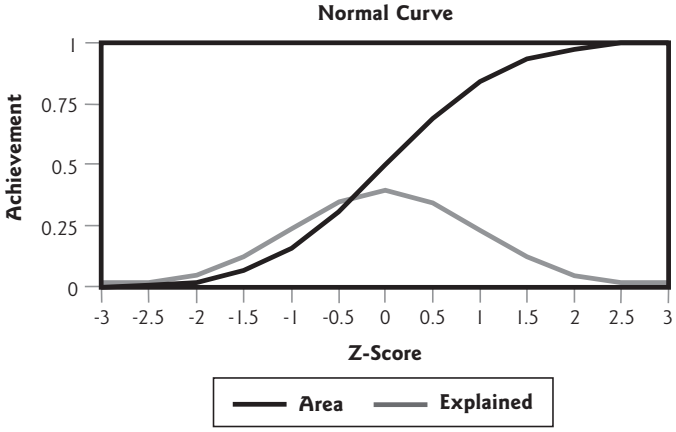
The PBA Model

The policy analysis model builds on the principles previously presented in the theory. Importantly, it is not an analytical model, such as regression, designed to estimate the magnitude of relationships. It is a mathematical structure purposefully designed to represent the most important characteristics of school achievement derived from productivity research and from state school data. The purpose of the model is to accurately predict the largest achievement gains based on changes in the most cost-effective policies. In other words, the model is structured to optimize achievement by selecting the most cost-effective policies. This section addresses the following five issues: Representing effect size; measuring effectiveness; predicting actual achievement; the importance of initial conditions; and predicting a change in achievement.

Representing effect size. A critical element of the PBA model is the function representing effect size—the magnitude of the relationship between the policy variables and achievement. Because there is a built-in ceiling to achievement tests, the relationship between achievement and the variables is nonlinear. The percent of variance explained, the R^2 from a regression equation, is the logical function. It can be estimated by means of statistical analysis, and it allows for an optimization process not workable with linear relationships. The relationship between the total and explained variance is depicted by the following illustration. The achievement distribution (Total) and the distribution explained by the policy variables (Explained) are represented by normal curves, with explained portion being a proportion of the total.[8] (See Figure 2.)

The normal curve of the explanatory variable is mathematically integrated (summed to find the area under the curve). Thus, the explanatory variable is measured in standard scores (Z-scores), and achievement is measured in percentiles (area under the normal curve). The following illustration depicts the relationships between the distribution of the explanatory variable, the integral of the explanatory variable, and the achievement variable. For any value of R^2, the normal curve can be transformed to an S-shaped curve.[9] (See Figure 3.)

Figure 3
Representations of Effect Size



Measuring effectiveness. Previously, several categories of policy variables were listed, and each category has constituent variables. Because the constituent variables are most likely correlated, it is impossible to precisely measure the unique and common contribution each variable makes to achievement; that is, the contribution a classroom teacher makes to a student's achievement cannot be precisely separated from the contribution a special reading teacher or a teacher's aide might make to his or her achievement. Importantly, every constituent variable also has an effectiveness component; that is, not all administrators, teachers, reading teachers, or aides operate with equal effectiveness. Again, the constituent variables within the categories are usually correlated, so it is impossible to precisely measure the contribution effectiveness makes to achievement. Nevertheless, it is possible to estimate the total contribution effectiveness makes to achievement across all categories.

It is possible based on factor theory to measure the total achievement contribution—common and unique—of the conceptually and statistically related variables within categories, more appropriately called factors. The constituent variables for the Minnesota data were combined into factors: Staff quantity; staff qualifications; instructional materials; and SES. When achievement was predicted based on these factors, there was sizeable error, i.e., the difference between the predicted achievement and the actual achievement (the residual) was fairly large. Was the error systematic or random over time? In other words, did some schools consistently produce higher (or lower) achievement than what was predicted? The answer is yes, i.e., a portion of the error is systematic. Over a number of years, some schools consistently did something positive to produce higher than expected achievement taking into consideration the resource factors and SES. Some schools did the opposite, consistently producing lower achievement. This tendency to produce (or not to produce) achievement is measured by averaging the school residual over time (fixed effect estimation). This unobserved indicator of achievement production has been labeled "effectiveness" and most likely consists of some form of organizational and instructional behavior as proposed by Levin and Walberg.

Predicting actual achievement: The importance of effectiveness. The only way to accurately predict actual achievement is by comparing schools within the same state using the same achievement and explanatory variable measures. From these data, effect sizes for

Table 1
Estimates of Effect Size for SES, Resources, and Effectiveness

Achievement	*SES*	*Resources*	*Effectiveness*	*Error*	*Sum*
Mathematics	0.550	0.035	0.340	0.075	1.00
Reading	0.620	0.090	0.230	0.060	1.00
Mean	0.585	0.063	0.285	0.068	1.00

resource factors, SES, and effectiveness are estimated. The following production function predicts actual achievement (AA) from the resource factors and SES, as well as the contribution made by effectiveness, with a margin of error:

$$AA = \sum R^2 * \text{Resource factor} + R^2 * SES + R^2 * \text{Effectiveness} + \text{Error} \quad (1)$$

If effect size estimates (R^2) for the resources are used from other studies and they are higher than those from the state database, these estimates will predict achievement levels higher than the actual achievement. In this case, the production function can only be balanced to equal the actual achievement by reducing the contribution of effectiveness. In other words, if smaller classes are thought to make a larger difference and that difference is not reflected in the calculations for actual achievement scores, then schools must be ineffective in utilizing the full benefits of the smaller classes. This is a critical point worth restating. Lower class size predicts achievement only if the lower class size is implemented effectively. If a school does not meet the achievement level predicted by the class size, the only explanation is that they are ineffective. Conversely, if a school exceeds the achievement level predicted by the class size, they must be more effective in the implementation. Effectiveness is inextricably related to achievement production!

Regarding the theory of gravity, we know there is such as thing as a Graviton because we can measure its influence even though we do not know how it works. Regarding achievement productivity, we know there is such a thing as effectiveness because we can measure its influence even though we do not know exactly how it works. The following model explores this question: What are the possible characteristics of effectiveness, and how can they be incorporated into policy analysis?

The estimated effect sizes of the factors, taken from the Minnesota data set, are presented in Table 1. The staff quantity, staff qualifications, and instructional materials are included under the "Resources" factor. Because the factors are measured in terms of the R^2, the sum of the factors must equal 1.00: If one factor is increased, another factor must be decreased. More importantly, if the effectiveness factor is not included, the error is increased.

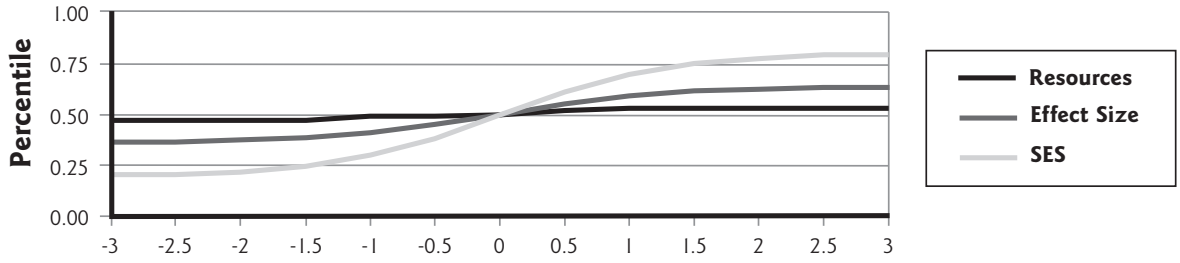
When plotted, the effect sizes appear as S-shaped curves with the height of the curve proportional to the effect size. Effect size is analogous to a hill, the steeper the hill the larger the benefit. As the effect size gets smaller, it approaches a straight line. (See Figure 4.) As will be discussed later, it requires energy (resources) to "climb the hill."

Table 1 and Figure 4 highlight the critical differences between this PBA paradigm model and other models of achievement productivity. In this paradigm, the nonlinear effect sizes are bounded because of the inherent floor and ceiling in achievement testing. The position on the S-shaped curve determines the marginal effect size unique for each school rather than a constant effect size common for all schools. Also, the influence of a policy variable cannot be estimated without taking into consideration the effectiveness of implementation.

Importance of initial conditions. Returning to the theory of gravity and the work of Galileo, an object continues to move in the same direction and at the same speed unless another force is applied. The original direction and speed are called the initial conditions. By knowing the initial conditions and the speed and direction of the intervening force, the new direction and speed can be calculated. Applying this principle to achievement production, any model must first accurately determine actual achievement based on the initial conditions before it can forecast a change of achievement based on the change of those conditions.

The current standings of the resource and SES variables are considered the initial conditions. These initial conditions are determined by a school's placement within the total population, as measured by Z-scores and percentiles; that is, the contribution to achievement made by any variable depends where on the curve the school is

Figure 4
A Representation of Effect Size



situated because the slope is always changing. Identifying the initial conditions for the effectiveness variables is addressed later.

Predicting a change in achievement. After the model accurately predicts actual achievement, it must be modified to accommodate the changes of policy variables, which will predict later achievement. A critical element of the PBA model is the function relating achievement with the various policy options. Because a change of variables likely requires a change of funding, a cost variable is added to the equation:

$$PA = \sum R^2 \$ f_{(z)} \quad (2)$$

where

PA = predicted achievement, and for every policy variable;
R^2 = estimated effect size;
\$ = incremental cost;
z = condition of the school on the policy variable;
and
$f_{(z)}$ is the nonlinear function representing the relationship between the policy variable and achievement.

A separate equation is constructed for each desired achievement outcome. The goal is to change various policy variables from their initial condition to their optimal condition to attain the highest potential gain in achievement, i.e., to change the value of Z. The change, or gain, in achievement is the difference between actual achievement (the old z) and predicted achievement (the new z).

Production Model Process: Steps for Implementation

The implementation of the model is divided into three broad steps: (1) Developing various policy options or scenarios, and simulating their influence on achievement, using estimates of effect sizes, estimated incremental costs, and the initial conditions of the policies; (2) evaluating the various scenarios based on the predicted achievement level; and (3) testing the success of the selected scenario through implementing the policy and measuring the accuracy of the prediction.

Developing policy options. The model evaluates achievement theories by simulating how various policies might impact achievement. Each combination of policy options is called a scenario.[10] The following resource and effectiveness factors with their constituent policy variables are available for inclusion in the simulation.[11]

- Resource variables—these variables, which are objects of funding, are identified in most state databases:
  - o Staff quantity, e.g., ratios of teachers, aides, instructional support, and administrators to pupils;
  - o Staff qualifications, e.g., education, experience, salary;[12]
  - o Instructional materials.[13]
- Effectiveness variables: There is no direct identification or measure of process variables in state databases, but an indirect measure of an effectiveness factor is available for every school and is of a substantial magnitude. The following characteristics are assumed to be the components of the effectiveness factor and are called effectiveness variables in the remainder of the paper.
  - o Instructional Effectiveness: Walberg identified these instructional characteristics—curriculum, method of instruction, instructional organization, home contribution, and time-on-task.
  - o Operational Effectiveness: Levin identified these operational characteristics—measurable outcomes, incentives linked to outcomes, productive technology.

Evaluating scenarios. After possible policy scenarios are developed, they can be evaluated via simulation to estimate the predicted gain in achievement. Those portions of the policy scenarios judged to be workable based on predicted achievement gain, cost effectiveness, and practical operational considerations are refined while the impractical portions are dropped from further consideration. This refining process is continued until a final scenario is selected for implementation. The following example provides more detail regarding this process.

Testing Scenarios. This model is theoretical as it has not been tested in an actual situation. If persuasive, it gives direction as to how the model could be implemented and the results tested. First, more research into the characteristics of an effective of curriculum and instruction program would be valuable, as well as research into the characteristics of organizational effectiveness. Second, the model does not represent a solitary circumstance; rather, it is a template over which any circumstance or condition can be constructed. In essence, it is not the model that would be implemented and tested; it would be an individual scenario describing specific conditions that would be tested. Each scenario describes a set of school policies and makes an estimate as to the associated achievement. The selected scenario is tested by way of a case study where the implementation of the selected policies is monitored and the accuracy of the predicted achievement measured.

The case study approach would determine if the hypothesized characteristics of the policy options are actually present and influential. If they are, the scenario is directly confirmed, and the model is indirectly corroborated. As more evidence is collected, the model can be enhanced. To put it another way, the theories of Walberg (curriculum and instruction effectiveness) and Levin (organizational effectiveness), as well as those of STAR[14] and class size reduction experiments can be tested simultaneously within the same model. The model actually poses this research question: Can a specific level of academic achievement be accurately predicted by implementing a specific set of policies?

Example of the Policy Analysis Model

Prior articles in this issue center on the nature of the relationship between policy options and student achievement, and on estimating the effect size of the relationship. The previous section of this article described the theoretical bases and the specifics of the policy analysis model. The previous concepts and estimates are now transformed into a practical policy analysis model. Let there be no doubt, there are no magical answers. The suggested method demonstrates the difficulty in identifying the underlying data and assumptions required for any thoughtful policy analysis. It is often said that research is only as good as the data. In the case of policymaking, decisions must be made without the benefit of perfect data. Therefore, good policy depends on good judgments. These judgments are based on clear and comprehensive assumptions regarding the operations of the enterprise: What are the goals to be accomplished; what policies will influence behaviors; and what behaviors will achieve the established goals?

To follow is a description of how a policy analysis model might work in seven steps, as follows: Optimization principles; school

Table 2a
School Profile

	Grade Level							Total	Cost ($)	
	K	1	2	3	4	5	6		Average	Total
Student Enrollment	40	40	40	40	40	40	40	280		
Number of Teachers	2	2	2	2	2	2	2	14	$60,000	$840,000
Pupil/Teacher Ratio	20	20	20	20	20	20	20	20		
Number of Aides	1	0	0	0	0	0	0	1	$30,000	$30,000
Support (Reading Teacher)								1	$70,000	$70,000
Administrator								1	$70,000	$70,000
Total Instructional Staff								17	$80,000	$1,020,000

Table 2b
Statewide Statistics for Staffing Ratios

Staff per 1,000 Students		Mean	Standard Deviation	Z-Score	Percentile
Teachers	50.00	67.97	13.28	-1.35	8.80
Aides	3.57	22.14	20.51	-0.91	18.26
Support Positions	3.57	3.77	1.93	-0.10	45.90
Administrators	3.57	2.90	1.56	0.43	66.65

profile; estimating effect sizes; determining the initial conditions; the optimization process; interpretation of results; and the policy development process. The description of the process is followed by a discussion of the value of a policy analysis simulation and other considerations.

Optimization Principles

It is possible through mathematical programming to optimize the policy alternatives; that is, to select the best combination of policy alternatives based on their effect sizes, incremental costs, and initial conditions. For the optimization, a set of simultaneous equations is developed, one equation for each desired outcome including all of the influential policy variables. Another equation is constructed to calculate the cost of increasing the level of the policy variables. It is also possible for some variables to be decreased and the cost to be reduced. The goal is to select the optimal level for each policy variable that produces the highest level for the combined achievement outcomes while staying within an established cost limit.[15]

School Profile

To illustrate the PBA model, a hypothetical school is profiled. In reality, the data would be entered for the school in question along with the necessary statewide data. The information includes the number of students and staff in the various grades; average and total salaries; and the statewide means and standard deviation for the staffing ratios (staff per 1,000 pupils). From this data, the Z-scores and Percentiles (Ptile) are calculated. (See Tables 2a and 2b.) Additional data would be added to the profile if they were to be incorporated into the policy analysis. The school profile defines the specific initial conditions necessary to predict a change in achievement.[16]

Estimating Effect Sizes

The preceding article discussed the process of estimating effect sizes and provided estimates from various sources. The estimates from the Minnesota data set are the estimates used for the resource variables in this example. For the effectiveness variables, the estimates are those derived from Walberg. Because the Walberg estimates are not from the Minnesota data set, it is reasonable to substitute different estimates. Because there is an estimate for the effect size of the entire effectiveness factor, the average for the constituent variables could be a starting point, with adjustments made based on the judgments of the policymakers.

Determining the Initial Conditions

The initial conditions reflect the position of the school on the respective variables as measured first in Z-scores and then percentiles. The initial conditions of the variables must be set so the predicting equation equals the actual achievement. There are three groups of variables: Resource variables; effectiveness variables, including a portion of the SES variable thought to be subject to some policy influence; and fixed variables outside the influence of policy—the other portion of SES and error.[17]

The initial conditions for the resource variables and SES are standardized measures from the state database. The initial conditions of the effectiveness variables are unknown but can be estimated. First, the school must judge the "quality" level for each of the variables. Because there is no standardized measurement scale, one must be devised. To match the method of measuring resources, the starting point of the scale is a Z-score of 0, with a standard deviation of 1. Based on this scale, each effectiveness variable is rated either up or down. These quality values (Q) also meet another condition; when

Figure 5
Setting Predicted Achievement to Actual Achievement by Adjusting the Initial Conditions for the Effectiveness Variables

	A	B	C	D	E	F	G	H	I	J	K		L
1	POLICY OPTION		INCREM		(R SQ)	(R SQ)		NOW			ACHIEVEMENT		
2		ADD	COST	TOTAL	READ	MATH	TOTAL	RATIO	Z	PTILE	READ	MATH	TOTAL
3	RESOURCES												
4	Teachers	0.00	60,000	0	0.058	0.048	14	50.00	-1.35	8.80	-2.39	-1.98	-4.37
5	Aides	0.00	30,000	0	0.035	0.035	1	3.57	-0.91	18.26	-1.11	-1.11	-2.22
6	Support	0.00	70,000	0	0.010	0.010	1	3.57	-0.10	45.90	-0.04	-0.04	-0.08
7	Admin	0.00	80,000	0	0.010	0.010	1	3.57	0.43	66.65	0.17	0.17	0.33
8	EFFECTIVENESS						Q	C	Z	PTILE	READ	MATH	TOTAL
9	Curriculum	0.00	5,000	0	0.047	0.047	0.50	0.198	0.70	75.74	1.21	1.21	2.42
10	Instruction	0.00	5,000	0	0.083	0.083	0.00	0.198	0.20	57.85	0.65	0.65	1.30
11	Organization	0.00	5,000	0	0.003	0.003	-0.50	0.198	-0.30	38.14	-0.04	-0.04	-0.07
12	Home	0.00	5,000	0	0.070	0.070	0.00	0.198	0.20	57.85	0.55	0.55	1.10
13	Time	0.00	10,000	0	0.051	0.051	0.00	0.198	0.20	57.85	0.40	0.40	0.80
14	Change SES	0.00	15,000	0	0.050	0.050	0.00	0.198	0.20	57.85	0.39	0.39	0.79
15	FIXED							0.198					
16	FIXED SES		N/A	N/A	0.500	0.500	0.00		0.00	50.00	0.00	0.00	0.00
17	ERROR		N/A	N/A	0.083	0.093	0.00		0.00	50.00	0.00	0.00	0.00
18			TOTAL	$0					SUM		-0.21	0.21	
19	Students	280	TARGET	$100,000	GAIN	0.00%			PREDICT		49.59	50.41	100.00
20	Per Pupil			$3,643		$0			ACTUAL		50.00	50.00	100.00

combined with the values of resource variables, they must predict actual achievement. To accomplish this, a parameter (C) is introduced which adjusts all the effectiveness variables, assuring that the equation equals actual achievement. This method answers the question: What initial conditions for the resource and effectiveness variables will predict actual achievement?[18] The initial condition for the error term is set to 0.

Actual achievement (*AA*) equals the sum of the resource variables (R) plus the sum of the effectiveness variables (E) adjusted by parameter (C), and the error:[19]

$$AA = \sum R_{(z)}R^2 + \sum E_{(z = Q+C)}R^2 + \text{Error}_{(z = 0)} \quad (3)$$

If the effectiveness variables were judged artificially high, the predicted achievement would be higher than the actual achievement. In essence, the parameter C becomes a "truth detector" for the quality judgments, and makes the appropriate adjustment. Actual achievement can be high only when the both the resource and effectiveness variables are at high levels. (See Figure 3.)

For total predicted achievement to equal total actual achievement, the initial condition parameter for the effectiveness variables (C) is .198.[20] (See Figure 5, Column H, Lines 8-14.) If actual achievement were higher than 100, the effectiveness parameter would increase, i.e., the school operations are more effective, and vice versa.

Optimization Process

The next step is to identify the most cost-effective policy options by automatically determining the best option through an optimization process. Many spreadsheet programs have an optimization feature. In Microsoft Excel, it is referred to as the "Solver." By identifying the target as the maximum gain in achievement, Solver will determine the best allocation of funds among the policy variables based on effect sizes, incremental costs, initial conditions, and an overall spending constraint.

In mathematical programming, the parts of the model are called the object function and the constraints. The object function is a mathematical function representing the goal to be attained, in this case the sum of various achievement measures. There are two types of constraints. The first type includes the mathematical functions representing the relationship between the various explanatory variables and the various outcomes. The second type includes the boundaries—maximums or minimums—for the variables. Importantly, there must be a boundary or upper limit to at least one variable, in this case cost, or there can be no end or conclusion to the calculations. Solver requires these parameters:

- Set Target Cell To:
  - The cell contains the object function or value to be attained, in this example the sum of the achievement measures.
- Equal To:
  - Maximum, minimum, or value. In this example, maximum is marked. The purpose is to find the values producing the maximum predicted achievement.
- By Changing Cells:
  - The range of cells is the values of the policy variables to be changed in order to obtain the maximum achievement level.
- Subject to the Constraints:
  - The maximum-, minimum-, or equal-to-values that reflect the assumptions regarding the school operations. Most importantly, the value of the additional cost must not exceed the predetermined value or target value. In this example, the constraints are set to prohibit any reduction of existing staff or a reduction in any of other policy variables.

Figure 6
Optimization of Policies

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	POLICY OPTION		INCREM		(R SQ)	(R SQ)	NOW				NEW			
2		ADD	COST	TOTAL	READ	MATH	TOTAL	RATIO	Z	PTILE	TOTAL	RATIO	Z	PTILE
3	RESOURCES													
4	Teachers	0.00	60,000	0	0.058	0.048	14	50.00	-1.35	8.80	14.00	50.00	-1.35	8.80
5	Aides	0.00	30,000	0	0.035	0.035	1	3.57	-0.91	18.26	1.00	3.57	-0.91	18.26
6	Support	0.00	70,000	0	0.010	0.010	1	3.57	-0.10	45.90	1.00	3.57	-0.10	45.90
7	Admin	0.00	80,000	0	0.010	0.010	1	3.57	0.43	66.65	1.00	3.57	0.43	66.65
8	EFFECTIVENESS						Q	C						
9	Curriculum	2.06	5,000	10,284	0.047	0.047	0.50	0.198	0.70	75.74			2.56	99.47
10	Instruction	2.77	5,000	13,851	0.083	0.083	0.00	0.198	0.20	57.85			2.77	99.72
11	Organization	1.52	5,000	7,583	0.003	0.003	-0.50	0.198	-0.30	38.13			1.02	84.54
12	Home	2.71	5,000	13,541	0.070	0.070	0.00	0.198	0.20	57.85			2.71	99.66
13	Time	2.31	10,000	23,052	0.051	0.051	0.00	0.198	0.20	57.85			2.31	98.94
14	Change SES	2.11	15,000	31,689	0.050	0.050	0.00	0.198	0.20	57.85			2.11	98.27
15	FIXED							0.198						
16	Fixed SES		N/A	N/A	0.500	0.500			0.00	50.00				
17			TOTAL COST	$100,000										
18	Students	280	TARGET	$100,000	GAIN	9.80%								
19	Per Pupil			$3,643		$357								

Figure 6 illustrates the optimization process. Solver changes the values in the ADD cells in column B, producing the highest gain in achievement while simultaneously taking into consideration the cost. To explain the process, the simultaneous elements are by necessity described sequentially.

The change of conditions and costs constraints. The heart of the simulation is displayed under ADD of the spreadsheet, which determines the new conditions of the policy variables producing the maximum increase in predicted achievement. The starting point of all variables is zero; therefore, a zero under ADD indicates no change in condition. An increase of the policy condition incurs a cost. This cost, which appears by variable under TOTAL (column D), was calculated by multiplying the values under ADD by those under INCREM COST (Incremental Cost) in column C. These are summed to reach a TOTAL COST of $100,00 (column D, line 17). The TOTAL COST is limited to a user-determined value or TARGET cost. For this example, the TARGET cost has been set at a $100,000 increase (column D, line 18). PER PUPIL indicates that expenditures are $3,643 per pupil (column D, line 19). This represents a GAIN of $357 per pupil, or a 9.8% increase. Based on the new policy conditions, the NEW levels are provided (columns K-N):

- TOTAL refers to resource variables, which is the number of teachers, aides, support personnel, and administrators;
- RATIO is staff per 100 students;
- Z refers to Z-score;
- PTILE refers to percentile.

The new Z-scores and percentiles are also provided for the effectiveness variables (columns M-N, lines 8-14). Note that when the percentile rankings for some variables move to a point of diminishing returns (≥90%), the other variables become more cost effective.

In this example, actual achievement for reading and mathematics is set at the mean, or 50th percentile, with a total of 100. Because the optimization is yet to take place, there are no values for the change from the initial conditions (ADD) or increased costs attributed to changing the initial conditions (TOTAL).

The change in predicted achievement. In simple terms, the optimization identifies the most cost-effective policy variable and increases the policy value to a point of diminishing returns, at which point it moves to the next most cost-effective variable. It moves through this sequence until the funding target is reached. At that point, the total achievement gain is at the maximum level.

The information regarding the achievement levels before and after the optimization is provided in Figure 7. For each of the policy variables and for each subject area, reading (READ) and mathematics

Table 3
Verification of Effect Sizes in Simulation

Variables	*R^2*	
	Reading	*Mathematics*
Resource	0.113	0.103
Effectiveness	0.254	0.254
SES	0.550	0.550
Total	0.917	0.907
Error	0.083	0.093
Grand Total	1.000	1.000

Figure 7
Achievement Gains through Optimization

	A	B	C	D	E	F	G	H	I	J	K
1	ACHIEVEMENT	READ	READ	READ	MATH	MATH	MATH		TOTAL	INCREASED	GAIN/
2		BEFORE	AFTER	GAIN	BEFORE	AFTER	GAIN		GAIN	COST	$10,000
3	RESOURCES										
4	Teachers	-2.39	-2.39	0.00	-1.98	-1.98	0.00		0.00	0.00	0.00
5	Aides	-1.11	-1.11	0.00	-1.11	-1.11	0.00		0.00	0.00	0.00
6	Support	-0.04	-0.04	0.00	-0.04	-0.04	0.00		0.00	0.00	0.00
7	Admin	0.17	0.17	0.00	0.17	0.17	0.00		0.00	0.00	0.00
8	EFFECTIVENESS										
9	Curriculum	1.21	2.33	1.12	1.21	2.33	1.12		2.23	10,284	0.22
10	Instruction	0.65	4.13	3.48	0.65	4.13	3.48		6.95	13,851	0.50
11	Organization	-0.04	0.10	0.14	-0.04	0.10	0.14		0.28	7,583	0.04
12	Home	0.55	3.48	2.93	0.55	3.48	2.93		5.85	13,541	0.43
13	Time	0.40	2.50	2.10	0.40	2.50	2.10		4.19	23,052	0.18
14	Change SES	0.39	2.41	2.02	0.39	2.41	2.02		4.04	31,689	0.13
15	FIXED										
16	Fixed SES	0.00	N/A	N/A	0.00	N/A	N/A		N/A		
17	TOTALS	-0.21	11.57	11.77	0.20	11.98	11.77		23.55	100,000	1.50
18	Optimized		61.57			61.98		123.55			
19	Predicted	49.79			50.20			100.00			
20	Actual	50.00			50.00			100.00			

(MATH), the BEFORE and AFTER achievement results are expressed as percentiles. The achievement gains for reading and mathematics are provided under READ GAIN (column D) and MATH GAIN (column G) respectively. These are summed under TOTAL GAIN (column I).

Based on the assumptions in this example, the predicted achievement gains due to the effectiveness variables (curriculum, instruction, organization, home, time, change in SES) as seen under TOTAL GAIN are positive, ranging from 0.28 to 6.95 percentile points. However, no gains are shown for resource (staffing) variables. All of the effectiveness variables would have to be at the point of diminishing returns before the resource variables would become cost-effective. The increased cost for each variable is found in column J. To assist in the evaluation, column K provides the results of cost-benefit analysis, giving the gain in predicted achievement for each policy variable based on an investment of $10,000 (GAIN/$10,000).

Verifying effect size. There is a running tabulation of the R^2 entered into the optimization model. In order to protect against the tendency to overestimate the influence of the policy variables, the sum is provided. (See Table 3.) These should and do sum to 1.00, including the error. These effect sizes correspond to those of the Minnesota analysis. It is important to start with a state database in order to establish some reasonable ranges for the effect sizes. As was pointed out earlier, having good SES indicators is critical in obtaining good estimates for the other factors.

The constituent variables should fit within the limits of the resource and effectiveness factors listed in Table 1. Remember, .05 was moved from the SES factor to the effectiveness factor for the previously stated reasons. Even with the resource variable in the simulation being higher than the factor from the data set, the resource variables are not as cost-efficient as the effectiveness variables. It is clear that if the effectiveness factor were omitted from the analysis, the error factor would be substantially larger and the predicted achievement much less accurate.

Interpretation of Results

If the results from this model are only as good as the assumptions, what are those assumptions? The PBA paradigm and model stand on two pillars: The relationship between achievement and the policy variables is nonlinear; and the most effective policy variables are those influencing a change in behavior. The degree of trust in the results from the PBA model is directly proportional to the commitment to these assumptions. Trust does not work in the reverse direction; that is, trust in the assumptions is not directly proportional to the commitment to the results. In other words, one must trust the results because the theory and model are persuasive rather than trust the theory and model because one likes the results. As the reader will soon see, the results from the PBA models are quite different than those from other models.

The critical parameters in the model are effect sizes, initial conditions, and incremental costs. Particular attention should be paid to the veracity of these parameters. The illustrative simulation identifies instruction as the best investment and the other effectiveness variables as the most cost-effective, but why? It is because the effect sizes for the effectiveness variables are larger than those for the resource variables, and the incremental costs are less. The estimates of the effect size for the effectiveness variables originated with Walberg and are supported by the analysis of the residuals, the fixed effects. The other element is the initial condition. The model assumes the initial condition for the effectiveness variables can be established by the judgments of policymakers. Just in case, they are adjusted by the effectiveness factor (C), so they are at least in the "ballpark." Clearly, this assumption must be tested.

The final element is the incremental costs. Could the incremental costs be wrong? Doubtful! While there is a certain amount of

guesswork in the other parameters, the incremental cost estimates should be far more accurate. There is an instructive "rule of thumb": If the incremental cost of one variable is double the incremental cost of another, then the effect size of the more costly variable must be double in order for the benefit of the two variables to be equal. In other words, incremental cost, the most accurate parameter, is the most influential. The model provides a potential gain per $1,000 calculation to show the relative potential of each variable. With the assumed initial costs, the effectiveness variables clearly have greater potential benefits. Remember, the potential benefits are tied to the initial conditions. If the initial conditions for the effectiveness variables are high, their potential benefits diminish, and the resource variables become cost-effective.

Clearly, the assumptions seeding the model are critical, and current research is not a source for exact answers. Nevertheless, the preponderance of evidence is in the direction of school effectiveness being a substantial determinant of achievement, and the model addresses this effectiveness by giving clues as to where to look. It must be stressed once again: This optimization model does not give a policy answer. In essence, it is a decision support system, or a calculation machine providing results based on the user-defined assumptions. While the optimization process will mathematically provide the best solution, the solutions may not be compatible with perceptions of the situation.

This being said, some broad principles do apply. Because the model is a simulation asking "what if" questions, the principles are in terms of "what if":

- What if the parameters in the illustration were valid?
  o The potential gain in achievement is substantial, most of which is associated with the effectiveness variables.
- What if the class size effect size is set to the value estimated from the STAR experiment (.1)?
  o There would be no change in the conclusion. The effectiveness variables are still more cost-effective. The effect size for the class size variable would still be smaller, and the incremental costs would be higher compared to the effectiveness variables.
- What if the actual achievement for the school were different?
  o Remember, the prediction formula must predict the actual achievement for the school in question. To achieve this equalization, an effectiveness factor (C) is introduced indicating how effective the school is. If the actual achievement is higher than predicted, then the school is more effective in implementation.[21]
- What if the target amount is changed?
  o As the target amount increases, so does the predicted achievement, but at a decelerating rate--the benefits gradually get smaller. Various predicted achievement levels for various funding targets: $50,000 = 20.78; $100,000 = 23.82; $150,000 = 24.75. As the school becomes more effective, the potential achievement gain diminished.

At first appearance, the model seems to treat each variable as being independent when in reality it is more likely that the variables work in combination. Achievement results are due to a combination of efforts, with resource and effectiveness policies working together. The staffing options can be effective only if clear directions regarding behaviors are provided. An obvious example is: If the goal is to improve music achievement, hire a music teacher and provide a clear set of expectations. While the illustration emphasizes the policies at the school level, surely district wide policies are also influential. In that vein, it is possible and maybe wise to have a highly skilled staff member provide service to more than one school building.

There are an infinite number of possibilities, so only the major points will be reported here. First, the incremental cost parameters are reasonably accurate, and the incremental costs for the effectiveness variables are most likely less than those for the resource variables. Second, changes in effect size and initial conditions must be substantial before there will be a change in the optimization results. Third, the resource variables become cost-effective only when the effectiveness variables are near the maximum, and that happens only when the actual achievement is substantially higher than the predicted achievement.

These results have consequences for the research reviewed in the earlier article in that it changes the research question. No longer are the questions, does class size make a difference, or how much of a difference does it make? The new question is: Under what set of policy and behavioral conditions does achievement improve, and by how much?

The Policy Development Process

Most importantly, the optimization model is an iterative process. Once the result for one set of policy options is developed, it must be evaluated and refined. If a particular set of policy options is unworkable, setting a variable constraint to a different level modifies outcomes. As policy options are narrowed, so is the target cost, bringing the policy analysis to a desired funding level.

In reality, the results are only as good as the assumptions, so at every step of the process the assumptions must be evaluated. In other words, the model is a tester of assumptions, or a tester of the relationships among policies, behaviors, and achievement. As such, the best policy scenario is most likely natural rather than unnatural, with a sense of beauty or elegance rather than complexity.

While Solver refers to the various policy options as scenarios, these are really various theories of achievement production. In some cases, there is research defining the characteristics and estimating the effect size, but in many cases the relationship between the policy, behavior, and achievement outcomes is common sense. Here is an illustration of an actual linkage between policy, behavior, and achievement. In the early 1970s when our daughter attended the Shaker Heights, Ohio school system, the board of education adopted a reading and writing policy applicable to all students, teachers, and families. Starting in the fourth grade, every student was required to read a book of their choosing every week and prepare a written summary based on a prescribed outline. The student's family was required to enforce the policy at home, inspect the written summary, and attest to its authenticity. Finally, teachers were required to review the summary and judge whether it met the prescribed standard. If not, the report had to be redone. Reading and writing achievement improved. No research study was required.

This example emphasizes a theory of time-on-task; that is, the more time spent on an activity, the greater the performance. This is a possible scenario for inclusion in a policy analysis optimization by estimating the effect size and incremental cost. There are many other possibilities too numerous to fully discuss here, but the work of Levin, Walberg, and those mentioned in earlier articles in this

issue are starting points. Each school will have to critically evaluate their performance and decide what are the most pressing issues to address. Again, there is no single solution to all problems.

While many people think SES is the best predictor of student achievement, this is not the case. The best predictor of achievement is whether the student received instruction in the subject matter included in the achievement test. Students who have had a class in algebra consistently perform better on an algebra test than students who did not. Unfortunately, data for the effectiveness variables are limited, and the shortcoming must be compensated for by stringent analysis. Educators with expertise in several specialties—curriculum and instruction, administration, finance, social foundations—should bring their expertise to bear in analyzing each possible scenario. In this search, each school must do its own critique, answering the following questions:

- What are the appropriate outcome goals?
- What are the best educational practices?
- Where does the school stand in relationship to best practices?
- Are there model schools to emulate?
- What policies will most influence the desired behaviors of instructional staff, students, families, and, when possible, the community?
- What is the process to assign and monitor behavior with regard to training, written clarification, individual assistance, progress reports, evaluation, and rewards for success?
- What financial resources are required for additional staff, the purchase of additional time from existing staff, instructional materials, and specialized facilities?
- What is the estimated effect size to be accrued from the implementation of the policy?
- What is the feasibility of an effective implementation?

After the possible policy scenarios are developed, they can be entered into the optimization model where alternatives are evaluated by estimating the respective potential achievement gains. Instead of relying on opinion or on a review of the research literature, this policy development model requires a clear and comprehensive statement of the alternatives followed by a critical and comparative evaluation of the alternatives based on cost and potential benefits.

Other Considerations:
General Principles of the Optimization Model

There are other techniques to make the model more sophisticated:

- It is possible and even desirable to set boundaries for the policy variables. The boundaries consist of maximum and minimum levels, which the optimization process cannot exceed.
- Boundaries can be set so that one variable with a positive slope can be limited in order that another variable can be increased.
- It is possible to include policy variables with negative slopes, measuring the potential gain from reducing costs in these areas and applying the funds to another more productive area. These are called opportunity costs.
- It is possible to include non-achievement goals in the model as long as there is a measure of attainment, a measure of the initial conditions, estimated costs, and estimated effect sizes.

Solver creates several reports to assist in the analysis of the scenario. The "Sensitivity Report" contains information demonstrating how sensitive a solution is to changes in the formulas used in the scenario. It measures the increase in the predicted achievement level for a unit change in each of the determinants and constraints. The "Answer Report" provides the predicted achievement level; the original and final values of the determinants; and information about the constraints. The "Limits Report" lists the achievement levels and the determinants with their values, and lower and upper limits.

Value of a Policy Analysis Simulation

Building a simulation model has several potential benefits:[22] The exercise of building a simulation model often reveals structures and relationships not previously apparent. As a result, there is a greater understanding of the complex process of achievement production. The modeling process can identify areas where additional research is needed. Having built a model, it is possible to analyze it mathematically to help suggest courses of action not otherwise apparent. Experimentation with many options is possible with a model whereas it is often not possible or desirable to experiment on the actual situation. Many policy options can be tested, first separating practical from impractical solutions. If a satisfactory policy option is identified during the simulation process, it gives clear directions as to how it could be implemented and tested in an actual situation. As more experience and knowledge is gained, the model is enhanced.

When decisions are made based on opinion, the underlying assumptions regarding policy actions, costs, and predicted benefits are mostly ambiguous; therefore, there is no method to test the likelihood of achieving the desired goals. While productivity research may give some helpful direction, research in and of itself does not provide sufficient information regarding particular situations (policy actions and costs) to accurately predict outcomes. Only through a comprehensive policy analysis model can the underlying assumptions be clearly stated, evaluated, and tested.

A Final Word

In the early 1900s, the notion of gravity took a major turn. Einstein developed his theories of general and special relativity based on the idea that space is actually curved—nonlinear. Years later, the theory was confirmed by experiment showing that light from distant stars indeed curves around the sun on the way to earth. Space travel is calculated by his equations. While not of the same magnitude, it is reasonable that the relationship between achievement and policy variables is better explained by a nonlinear function, and it is worthy to test by experiment. After all, there are no experiments demonstrating that the relation is linear!

Admittedly an exaggeration, here is a characterization between the effective and noneffective method of allocating of resources. This first is called the "Professor Henry Hill" method after the lead character in the Meredith Wilson musical, "Music Man." Hill, a traveling salesman, convinced the people of River City to purchase from him bright new uniforms with shiny buttons for the school band, and in return he could make beautiful music solving all the "troubles here in River City." Once he got the money, he employed the "think method" of instruction. If the students would "think" how nice it would be to march down the street in their magnificent

uniforms with their parents and community cheering them on, they would be able to skillfully play their instruments. Sure enough, it worked and everyone was treated to a magnificent parade with "Seventy-six Trombones."

The second example is called the "Carnegie Hall" method after a common musician's joke. While walking down the streets of New York City, a person asked a stranger, "How do you get to Carnegie Hall?" The stranger replied, "Practice, practice, practice." Imagine a situation where students are in an instrumental music class learning to play an instrument. They meet regularly, receive structured and competent instruction, take their instrument home, and the parents oversee thirty minutes of practice every day. At each step, there is a clear policy directing student behavior. It does not take a sophisticated research study to determine the difference of musical quality being produced by the two paradigms.

References

Achilles, C.M., B.A. Nye, J.B. Zaharias, and B.D. Fulton. "The Lasting Benefits Study (LBS) in Grades 4 and 5 (1990–1991): A Legacy from Tennessee's Four-year (K–3) Class-size Study (1985–1989)." Project STAR. Paper presented at the North Carolina Association for Research in Education, Greensboro, North Carolina, January 14, 1993.

Barnett, Raymond A., and Michael R. Ziegler. *College Mathematics for Management, Life, and Social Sciences*. San Francisco, CA: Dellen Publishing Company, 1984.

Coleman, James S., Ernest Q. Campbell, Carol J. Hobson, James McPartland, Alexander M. Mood, Frederic D. Weinfeld, and Robert L. York. *Equality of Educational Opportunity*. Washington, DC: U.S. Department of Heath, Education, and Welfare, Office of Education, 1966.

Feynman, Richard P. *The Character of Physical Law*. Cambridge, MA: The MIT Press, 1965.

Hedges, Larry V., Richard D. Laine, and Rob Greenwald. "Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes." *Educational Researcher* 23 (April 1994): 5-14.

Kuhn, Thomas S. *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press, 1970.

Levin, Henry M., "Raising School Productivity: An X-Efficiency Approach." *Economics of Education Review* 16 (June 1997): 303-311.

Phelps, James L. "Optimizing Educational Resources: A Paradigm for the Pursuit of Educational Productivity," *Educational Considerations* 35(Spring 2008): 3-18.

Phelps, James L. "Measuring and Reporting School and District Effectiveness." *Educational Considerations* 36(Spring, 2009): 40-52.

Walberg, Herbert J. "Improving the Productivity of America's Schools." *Educational Leadership* 41 (May 1984): 19-27.

Williams, Hilary P. *Model Building in Mathematical Programming*. 2nd ed. New York: John Wiley, 1985.

Endnotes

[1] In particular, the previous article, "A Practical Method of Policy Analysis by Estimating Effect Size," led to a number of underlying assumptions that will guide the analysis here. See Appendix A for a list of these.

[2] The current achievement theories and models tend to follow the interpretation of the physical science laws: If one variable changes, the consequences are automatic. If students leave the classroom, does the knowledge of the remaining students increase automatically and at the speed of light? Do teachers and students, like "mother nature," automatically know what to do, or must another process take place?

[3] All subsequent references to Glass and Smith in this article refer to Gene V. Glass and Mary Lee Smith, *Meta-Analysis of Research on the Relationship of Class-Size and Achievement* (San Francisco, CA: Far West Laboratory for Educational Research and Development, 1978).

[4] All subsequent references to Hedges et al. in this article refer to Larry V. Hedges, Richard D. Laine, and Rob Greenwald, "Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes," *Educational Researcher* 23 (April 1994): 5-14.

[5] Correspondingly, a substantial number of research studies openly state a purpose of proving Eric Hanushek, a critic of these types of studies, wrong! The same was true in the 1970s when Coleman et al. (1966) issued the report, *Equality of Educational Opportunity*, with a conclusion showing the substantial relationship between achievement and socioeconomic status and a smaller relationship with resources.

[6] All subsequent references to Levin in this article refer to Henry M. Levin, "Raising School Productivity: An X-Efficiency Approach," *Economics of Education Review* 16 (June 1997): 303-311.

[7] All subsequent references to Walberg in this article refer to Herbert J. Walberg, "Improving the Productivity of America's Schools," *Educational Leadership* 41 (May 1984): 19-27.

[8] There is also an error distribution, or residual, not shown.

[9] Notice the similarity in shape between the integral of the normal curve and the "learning curve."

[10] "Scenario" is the description used in the Microsoft software, to be discussed later.

[11] Any policy variable can be included in a scenario if the effect sizes and incremental costs can be estimated.

[12] Available, but not included in the illustration because of small effect size estimates and limited space.

[13] Available, but not included in the illustration because of small effect size estimates and limited space.

[14] See Achilles et al. (1993).

[15] The details are provided in Phelps (2008).

[16] While necessary for this policy analysis by policymakers and practitioners, reporting the status and progress of schools to the public is valuable as well. A comprehensive review of these issues is available in Phelps (2009).

[17] The staffing qualifications and instructional materials categories are omitted from the illustration because of limited space and their small effect sizes, but they could be included as resource variables in a full simulation. The organizational effectiveness category is also omitted because there are no estimates of effect size.

[18] The Z-scores are converted into percentiles, and the predicted achievement equation is made to equal actual achievement by determining the value of C.

[19] See Appendix B.

[20] The value of C is derived via Microsoft Excel Solver. The Target Cell is set to 100 (the Actual Achievement level), By CHANGING CELLS is the value of C.

[21] Various actual achievement values were entered with the corresponding C values: 80 = -.60; 100 = .26; 120 = 1.5.

[22] Hilary P. Williams, *Model Building in Mathematical Programming*, 2nd ed. (New York: John Wiley, 1985), 3.

Appendix A
Underlying Assumptions for the Policy Analysis Model

- The teacher/pupil ratio is a more appropriate policy measure of teacher concentration than is class size (pupil/teacher ratio).
- Influence of SES is critical in measuring the effect size of the teacher/pupil ratio.
- The evidence from the previous articles in this issue discounts the Glass and Smith proposition of increased marginal gains for class sizes under 15, so their proposition will not be included.
- The R2, a nonlinear measure of effect size, has distinct advantages over the other options for developing a comprehensive policy strategy.
- There is substantial collinearity among most educational variables and the estimated effect sizes depend on the attribution of the common variance. The effect size estimate varies depending on how the common variance is attributed. Therefore, a maximum to minimum range is an appropriate estimate.
- Because of the substantial collinearity, it is best to combine the instructional variables into conceptual and statistical categories and estimate the effect size of the entire category.
- It is likely that the instructional and organizational variables work cooperatively with the resource variables.
- Some schools are more effective in implementing the policy options. If more attention is paid to the implementation, it is possible to achieve more than the predicted gain based on resource level alone.

Appendix B
Logistic Growth Curve and Calculation Formulaes

Logistic Growth Curve

Logistic growth: Rate of growth is proportional to the amount present and to the difference between the amount present and a fixed amount (Barnett and Ziegler 1984, 819).

$dy/dt = ky(M-y)$ with $k, t > 0$

where

k= rate

M = maximum

or

$y = M / 1 + ce-Mt$

Calculation Formulas

In Cartesian geometry, the origin of a graph is the intersection of the X- and Y-axes. This is the case with standard or Z-scores at point X = Zero and Y = Zero. The origin of the graph changes when Z-scores are transformed into percentiles. Because the mean (50th percentile) of the explanatory variable is equal to the mean of the achievement variable, the origin of the percentile graph is at the 50th percentile; and because the normal curve is symmetrical above and below this point, half of the distribution is above, and half is below. Finally, when the explanatory variable is a zero Z-score or the coefficient is zero, then the achievement variable is at the mean or 50th percentile.

Achievement is calculated from the percentile position of the school and the effect size, the R^2. The initial condition determines the percentile position for the actual achievement and the optimal condition for the predicted achievement.

- The contribution each variable makes to achievement is calculated from the percentile position and the R^2. The percentile position is calculated from the initial or optimal condition Z-score by the Excel function, NORMSDIST:

  Percentile = NORMSDIST (z)

- Because a policy variable at the mean predicts achievement at the mean, the calculations are the contributions to achievement above or below the 50th percentile.
- To calculate the contribution (the difference from the mean), .50 is subtracted from the percentile and multiplied by the R^2:

  $\Delta = (Ptile -.5) * R^2$

- The contributions made by the variables, the Δ's, are summed. Because these are measures above and below the mean, .50 must be added to the sum of the individual contributions to obtain the predicted achievement level:

  $PA = \sum \Delta + .50$

Closing Essay: A Journey, Not a Destination

James L. Phelps

Much of the motivation and ideas for the articles in this special issue originated with my dear friend, Maris Abolins, Professor Emeritus of Physics at Michigan State University. We started as neighbors and, as our kids grew up together, we socialized frequently. He is responsible for my interest in physics. I would read a physics book, which would become the subject of our next dinner conversation (while our wives talked about other, more social topics). Instead of a compilation of facts, physics became a way of thinking about problem solving. The "unified field" theory was the start of my new thinking. There are four fundamental forces in nature: The strong force holding the atom together; the weak force dealing with the decay of the atom; electromagnetism; and gravity. Subatomic particles are responsible for these forces. Einstein tried to combine these four forces into one comprehensive theory, but there was insufficient experimental information to be successful. While some of the forces have been united into a theory (relativity and electrodynamics by American physicist Richard Feynman), gravity remains illusive. Was it possible to unify the various aspects of achievement production into a comprehensive theory? I wanted to give it a go! A unified theory might provide ideas helpful for improving research; professional training and practice; and, therefore, student achievement.

The individual pieces of a unified achievement production theory were scattered about, but I had not taken the time to assemble them. According to Glass and Smith (1978), relationships might not be linear, which started my thinking.[1] There were some efforts in the field of mathematical programming, e.g., data envelopment analysis (Silkman 1986), but after investigating these I found them wanting. "Fixed effect" analysis was in the economics literature, but the idea that it represents educational effectiveness had not been fully developed. Again, there were possibilities. Cost-effectiveness was addressed more substantially by Levin (1988), but not in a way to influence policy decisions. There were large controlled experiments, but the emphasis was on class size and not on a wide range of potentially influential variables. Little attention was paid to how several variables might work together. Economists were largely in the forefront of research, and there was little integration of the instructional and organizational aspects as suggested by Walberg (1984)[2] and Levin (1997). There is a great deal of ambiguity as to the purpose and conclusions of research. The research seems to be divided between what advocates more resources and what advocates organizational changes in order to improve education. There is little discussion regarding how they might work together. I wanted to rethink the fundamentals and see if these scattered pieces could be combined in some meaningful way.

After a professional meeting where the idea of simultaneous equations was raised, I started by writing down a number of basic equations to see if I could find some uniting principles. When each of the equations was graphed,[3] there were straight lines going every which way. There was no rational way to unite or choose among the alternatives. The only interpretation was to provide unlimited resources for all variables with positive slopes, hardly a practical or unifying strategy. With enough money, all schools could get perfect scores, a doubtful outcome. And what would be done with the variables with negative slopes—eliminate them all together? There was no practical method of evaluating alternatives. There were logical contradictions among the pieces. Instead of clarity, the exercise caused anxiety and confusion.

What made Albert Einstein so unique was his willingness to take on problems characterized by contradictions between explanations and experimental evidence. His contributions were monumental because he was able to make sense out of those contradictions. Richard Feynman was also a maverick in much the same way. In his books, Feynman writes about returning to the "first principle" when tackling intractable problems. He would start with the first principles—the basic principles underlying the phenomenon. He would test these principles to determine if they could stand strict scrutiny. If not, he would replace questionable principles with better alternatives. With the new principles in place, new solutions evolved. In essence:

- Flawed first principles lead to contradictory explanations and inaccurate predictions.
- Superior first principles lead to improved explanations and more accurate predictions.

Reviewing the productivity research is a strenuous exercise, as demonstrated by the earlier articles. Even the most diligent and ardent observer of achievement productivity research will have difficulty in reaching meaningful conclusions. There is "something for everyone." There is least one study supporting every possible policy conclusion. As a result, research has little value in solving everyday problems. It raises the question: Why conduct further research if the inevitable conclusion is the same—every option is effective!

There is no set of rules consistently and effectively applied to the many diverse educational situations. Instead, there are different and conflicting rules applied universally, discounting the unique situations. What are those "achievement rules"? The "Glass Rule" is to lower class size to one even though there is not enough money to do so. The "Hanushek Rule" is reduction of class size sometimes

---

James L. Phelps holds a Ph.D. from the University of Michigan in Educational Administration. He served as Special Assistant to Governor William Milliken of Michigan and Deputy Superintendent in the Michigan Department of Education. Active in the American Education Finance Association, he served on the Board of Directors and as President. Since retirement, he spends a great deal of time devoted to music, composing and arranging, playing string bass in orchestras and chamber groups, as well as singing in two choirs. He resides with his wife, Julie, in East Lansing, Michigan.

works and sometimes does not work; it all depends. The "Hedges Rule" (Hedges, Laine, and Greenwald 1994) is not to spend money on reducing class size, but spend money on whatever local decision-makers think is important. The "Tennessee Rule" (Achilles et al. 1993) is to lower class size. The "California Rule" (Bohrenstedt and Stecher 1999; 200s) is not to lower class size. The "Walberg Rule" is to change the curriculum and instructional programs, but with little direction as to how much and under what circumstances. The "Levin Rule" (Levin 1997) is to select the most cost-efficient programs, but by how much and under what circumstances? There was one common scheme. Every positive result reached the same conclusion: Increase funding without limit. Clearly, contradictory conclusions proliferate in achievement production research!

These "rules" are by-products of partial models; there is no single paradigm or comprehensive model encompassing the various aspects of the partial models.

The "reduce class size" or "spend more" rules are neither paradigms nor well-specified theories to test. Nevertheless, each piece of research has value in that it is a piece of a complicated puzzle. But the pieces have not yet been assembled into a mosaic for a clear image to appear. This is not to criticize the research as being bad. It points out the problem of reaching meaningful conclusions from research which has fundamentally different assumptions. What is missing is a set of first principles based on logic and evidence; and how the principles complement each other, and how accurately they explain and predict the phenomenon.

It is not possible to have multiple explanations for the same phenomenon—although it is possible to have several theories. After thorough testing, there must be just one theory which best explains and predicts the phenomenon. One of the basic assumptions of physics is that the physical laws apply everywhere in the universe. (It is science fiction when scientists apply different, untested laws.) Science is the pursuit of the best explanation with the best predictions. Regarding the explanation, the same laws apply in every situation, but when circumstances vary the solutions must vary. There cannot be identical solutions for varying circumstances. The influence of class size or any other variable must be the same in classrooms with similar conditions or it would be impossible to conduct research and to formulate explanations. Without this assumption, achievement production is reduced to opinion, with every opinion having equal, but not explanatory or predictive, value. But when school circumstances are different, there must be different solutions. The review of the achievement production research is abundant with contradictions regarding the statistical significance, shape of the relationships, effect sizes, and even the major determinants of achievement. Therefore, each piece of research produces a different explanation but the same solution, "unlimited more." I started to think in terms of some basic concepts, as follows: (1) *Similar circumstances must produce similar results; and there can be only one set of laws best explaining and predicting those results;* and (2) *Within the laws, different circumstances (parameters) must produce different solutions*. The challenge is to define the applicable laws and the influential circumstances.

Why the Contradictions?

Achievement research mostly relies on statistical models, which do not necessarily represent achievement production. Statistical models, in and of themselves, do not represent unified and coherent assumptions in all situations; they are tools to estimate the probabilities of relationships. Moreover, statistical models are not representations of the "real world." Rather, they are more like calculation machines providing a set of numbers in response to input numbers and instructions provided by the researcher. If the input numbers are good and the instructions are good, the conclusion might be good. Most importantly, the conclusions are not automatically good just because, "The model said so!"

Over time, statistical models have tended to become the mathematical representation of achievement production. In other words, the statistical models now de facto determine the first principles without further consideration of more appropriate principles. What is the first principle inherent in statistical models? The relationship between achievement and all explanatory variables is linear, so more of any explanatory variable will produce more achievement without limit. This principle is a primary source of the contradictions.

Should the researcher trust the conclusions and accept the model or trust the model and accept the conclusions? Can the conclusions be critiqued without fully critiquing the assumptions? Perhaps there is too much trust in the principles inherent in the statistical models and too much acceptance of the conclusions.

In many cases in the natural and behavioral sciences (gravity and the "learning curve," for example), mathematical representations were outgrowths of observations and possible explanations (theories). Only after the mathematical representation is developed are the predictions tested. In statistical analysis, the process is combined; the statistical model is the explanation (theory), the mathematical representation, and the testing mechanism. There is little questioning if the statistical model accurately represents the situation. As soon as the decision is made to use regression analysis, there is no further questioning if the relationships are nonlinear. Virtually all production function studies use regression analysis with the linear relationship principle. There is no follow-up to test the predictions, and the regression results are deemed to be reality. There is ample rationale and evidence to suggest that the achievement relationships are not linear and that nonlinear models should be considered. This is not to disparage these previous works. Without their efforts, it would be impossible to build something new.

There are reasons why a comprehensive, coherent, and unified modeling and testing process can be applied to achievement production. The purpose of this article is to identify those reasons. Are the proposed reasons perfect? No. Are they clear, comprehensive, unified, and coherent? Others will decide. It is not sufficient, however, to merely challenge the principles made herein; it is necessary to replace the principles with those better explaining achievement production and more accurately predicting achievement.

While overstated, there is an underlying truth to the saying: "If you keep on doing what you're doing, you will keep on getting what you're getting." If the same achievement production research is continued, the same conclusions will inevitably result. There seems to be sequence in bringing about change in what Kuhn (1970) calls "normal science."[4] First, there must be a new set of unifying and coherent principles, which become the basis of

research. The purpose of the research is to verify the principles. Once the principles are verified, they are used to train people who choose to apply these principles as a part of their profession. If the principles are correct, the research carefully conducted, the training effective, and the professional practice successful, the results will be rewarding.

Proposed First Principles

A set of first principles is proposed to address the contradictions associated with achievement production. The details and rationale for these principles are in the earlier articles. Here they are summarized in a different context, to be a foundation for future research, professional training, and practice.

These first principles were not conceived all at once. When I discovered what I thought was an inconsistency, I looked to a different knowledge base for possible answers. In essence, I was on a journey, which I briefly describe as a part of the first principles. You, the reader, are invited to retrace the journey, in the event you might discover another path.

Principle 1: Nonlinear Relationships

What started my analytical journey was the realization that achievement testing, like light, has its own "speed limit"—a perfect score—and as a consequence, the mathematical relationship between achievement and class size cannot be linear. Most certainly, it cannot be the curve suggested by Glass and Smith. The mathematical functions representing the theory of relativity are based on the idea that one can get closer and closer to the speed of light but can never exceed it. By demonstrating the mathematical difficulties in the Glass and Smith proposition, new thoughts came to mind regarding the nature of the determinants of achievement—the relationships must be nonlinear because there is a test ceiling and floor, and most likely the curve has a maximum and minimum (asymptotic at the top and bottom).

Years ago I heard a talk (I unfortunately do not recall where, or when, or by whom) about providing textbooks to classrooms in poorer parts of Africa. The speaker was raising the question, was it necessary for every student to have his or her own book? He concluded that it was not necessary. Students could share books and by doing so it was possible to save the expense and purchase books in other subjects. He drew a curve estimating the benefits of the number of textbooks—a diminishing returns curve. Ever since that talk, I have tried to identify circumstances where "more resources" do not eventually lead to diminishing returns. I have not identified any. It was important for me to know something about the research on learning, especially the "learning curves." Indeed, there is empirical evidence for a "learning curve," flat at the top and bottom.

By accepting the principle of nonlinear relationships, there are corollary principles.

- Every school has unique circumstances, identified by different points on the curves, meaning there is a different solution for every school rather than a single solution for all schools (principle of regression).
- There is a point where there become diminishing returns for all explanatory variables, rather than constant returns (principle of regression).
- There is an optimal point on each curve, allowing curves to be compared.

By changing one principle from linearity to nonlinearity, many of the contradictions were addressed.

Principle 2: Consistency of Components and Uncertainty of Measurement

In an publication using fixed effects analysis, I obtained a different set of explanatory variables for each year of data (Addonizio and Phelps 2006). There was no reason why the regression results should vary so much year to year. Then I realized slight changes in the correlation matrix would produce substantial changes in the order of significant variables in the step-wise regression results. As a result, the coefficient varied widely year to year. Simply put, basic laws cannot change year-to-year (if they could change by year, they could change by month, day, hour, or minute).

There were too many variables, and they were correlated. Merely entering all possible explanatory variables into a regression equation was not satisfactory; there was no theory driving the decision. The data were collected in categories: Staffing quantity; staffing qualifications; instructional materials; and proxies for socioeconomic status (SES). Rather than all variables working independently, it made more sense to have them working together; e.g., all staff work toward a common goal of achievement. The variables in each of the categories were used as explanatory variables against the various achievement measures. Averaging the coefficients over time addresses the time consistency of variables and consistency of coefficients issues. More importantly, the method represented a better explanation--conceptually similar and statistically correlated variables work together, not individually.

There was a second issue: The coefficient between achievement and an explanatory variable provides one estimate of the relations, but when a second explanatory variable is added, the results change. According to factor theory, two explanatory variables each make a unique contribution as well as a common or shared contribution. In essence, the contribution of any combination of correlated variables cannot be precisely measured. As is the case in quantum mechanics, there is inherent uncertainty of measurement. To deal with this uncertainty, the conceptually similar variables were grouped into factors and transformed indices by combining all the unique and common variance into the index. This provided an estimate of the contribution of the factor and upper and lower limits for each of the component variables.

Then there was the realization that educational research did not have an all-encompassing theory describing how all the various components fit together in a measurable and predictable way. Research mostly focuses on the pieces and not on the whole. Studies using different variables will undoubtedly get different results. Studies using the same variables get different results in different years. In order to estimate the basic laws:

- The basic laws must be comprised of the same explanatory variables although the coefficients can be different depending on grade and subject.
- Conceptually and statistically related variables must be combined in such a way to estimate the contribution of the variables within the group, and thus boundaries for the individual components.
- The coefficients of the basic laws are best estimated by averaging over time.

These principles are not a matter of personal preference; rather, they are a matter of statistical necessity. They explain some of the contradictions in the research--different variables and measures were used.

Principle 3: Accurately Representing Achievement Production

Education pursues multiple goals simultaneously. As a consequence, a single equation is not an accurate representation of the achievement production process, and a different formulation is required.

First, the achievement production system must be represented by simultaneous equations. There must be a separate equation for each achievement outcome and a way to control the cost of each of the variables, again in separate equations. This conclusion directed me to the field of mathematical programming, especially the books by Williams (1985) and Schrage (1991). None of the linear programming models worked because achievement was nonlinear (Principle 1). Was there a function representing the achievement/variable relationship that could be measured through some statistical process and could be solved using simultaneous equations? This became another dinner conversation, and Maris Abolins gave me *An Introduction to Error Analysis* (Taylor 1982). For the first time, I started to understand the reasoning behind the normal curve. I realized that the integral of the normal curve was the appropriate nonlinear function that could be measured by statistical regression. (It has a similar shape to the "learning curve" I was reviewing in another book. Both have the upper and lower limit properties.) All I had to do was find a way to formulate the necessary equations and solve nonlinear simultaneous equations. Back to mathematical programming I went and soon found software capable of accomplishing the task. Earlier software was cumbersome, but Microsoft Excel was easily available and easy to use.

Achievement production must be represented by a set of simultaneous equations representing each goal to be achieved, and must include an equation representing the costs. This addresses some of the contradictions.

Principle 4: Effectiveness Is An Integral Part of Achievement Production

I returned to Taylor (1982) and took note of the section dealing with systematic and random error. As a golfer, I immediately realized my hitting the ball consistently to the right was not random error, it was systematic. Systematic error can be separated from random. I had to correct my systematic error to improve my game. Now my topic became "fixed effect estimation' in econometrics. Because of my role in the Michigan Department of Education dealing with reporting school progress, I wrote the paper, "Measuring and Reporting School and District Effectiveness," (1988) building on my thoughts regarding factor theory and fixed effects. To borrow from my golf swing analogy, schools must correct their "slice" in order to improve student achievement. Including the notion of effectiveness in the simultaneous equations addresses some of the contradictions.

Principle 5: Achievement is derived from behavior

Again, the "eureka" moment came from reading physics, this time about gravity. The discussion was, how long would it take for the effects of the sun's collapse to reach earth? The answer is: At the speed of light. How long will it take for a change of class size to improve achievement? Surely, not at the speed of light. Actually, the change would not even be guaranteed. A change in achievement cannot be related to the number of students in the rooms, it must be related to the behaviors of the teacher, students, and parents. Somehow, the notion of behavior must be incorporated into the explanation and model. This notion explains some of the contradictions in research where the assumption of the regression model is that change is automatic.

Principle 6: Policies and Incentives Influence Behavior

The realization of the effectiveness and behavior notions brought new insights into my appreciation of the work of Walberg and Levin. Simply put, their ideas combined to make a plausible explanation. Policies influence behavior, and behavior influences achievement. In other words, their ideas were the reasonable explanations for the mysterious unobserved fixed effects or effectiveness. Even though there is much more research to be conducted in these areas, they do deal with some of the contradictions.

Principle 7: Policies Are Subject To Cost Constraints

Levin's influence on my thinking was substantial; cost-effectiveness must be included in any explanation of achievement production. With the simultaneous equation formulation, this was easily accommodated. This was the final piece of the puzzle and addresses what is perhaps the biggest incongruity in the regression formulation; that is, it is a basic inconsistency to advocate more of everything where there are fiscal constraints.

I have tried to carefully articulate the first principles in order for the reader to have the full context on which to critique the model.

Implications for Research

Are these principles valid? More accurately, are theses principles generally accepted as explaining achievement production? These principles are intended to be a beginning, not an end. It is important for there be a comprehensive discussion among those who are interested in the topic of achievement production in which they express their views and suggest improvements. As consensus is gained on the principles, attention can then be direct to research, training, and practice.

Are the opposite principles false? In most cases, each of the principles can be expressed in the negative, e.g., the relationship cannot be nonlinear and must be linear. By doing so, the distinctions are sharpened making the analytic process clearer.

Are these principles the foundation of current research, training, and practice? This is highly unlikely. There is little in the research literature regarding comprehensive theory; attention is mostly on specific issues. If I would identify the major weakness of research, it is the lack of consensus regarding the components of the underlying theory. After all, science is the testing of comprehensive theory, not the testing of unrelated assumptions.

Could these principles form the foundation of a new paradigm? Obviously, I think this is the case; it is why I have devoted my time and energies to this project. I wonder if others share this observation?

Could the new paradigm constitute the foundation of a normal science? My experience in academia and in the Michigan Department of Education leads me to believe that the pursuit of achievement excellence is not a scientific matter—it is mostly political. More emphasis is placed on more money and who gets the money

than how the money is used to improve the performance of students. Old research methods are repeated in hopes that they will miraculously produce different results.

If these principles are the foundation of the normal science of achievement production, will the practitioners of this normal science adhere to these principles? Schools of education are at a crossroad: Are they a branch of political science where opinion and perceptions are key, or will they move more toward normal science where theory, experimentation, and evidence are key?

As previously noted, the achievement paradigm must be thoroughly tested. First, individual profiles would be established for each school describing their unique situation regarding their standing on resources, SES, and effectiveness. Second, based on this information, the school would be asked to develop a set of policies and evaluate them based on the paradigm model and the predicted gain in achievement, and then select one for implementation. Third, they would implement the policies and collect information regarding the implementation. Finally, the information would be analyzed along with the actual achievement results to identify any relationships. Surely, such a planning process could do no harm. In contrast to the controlled class size experiments, such a regimen would provide a great deal of information upon which to address some of the unanswered questions:

- Do school organizations respond to policy changes, i.e., can good policies change the behavior of the instructional staff?
- What are the successful policies and effective implementation strategies?
- How does a change in instructional staff behavior influence a change in student, family, and community behavior?
- Can school policies influence family and community behavior?
- How are the changes in behavior translated into higher achievement?

Implications for Professional Preparation

Forrester (1980, 11) had some perceptive and instructional observations regarding organizations directly applicable to education:[5]

> For the most part, and in spite of lip service to the contrary, managers are usually decision-makers, not policy makers. The distinction is crucial. People can make decisions without knowing why. Decisions tend to be capricious and are dominated by short-terms pressures. A decision-maker runs an organization, but a policy-maker designs an organization. The distinction is like that between an airplane pilot and the airplane designer. It is the challenge of the designer to create a system that can function as intended in the hands of the kinds of operators who will be available. Seldom are school systems designed. We know that aircraft must be skillfully designed to operate properly, but the same attitude has not yet been generally extended to the much greater complexity of a school system. Here is the challenge and the opportunity for the teaching of management policy—teaching the design of the school systems rather than piloting. Modeling can provide the process for shifting the more responsible levels of management from being school system pilots to school system designers—to shift from coping with day-to-day crises to creating a social system that can be run by ordinary people without continuously recurring crises.

Actually there are many specialized people involved in airplane design: aeronautical, mechanical, and electrical engineers, to name a few. They work together in building a sophisticated product because they were trained within a common scientific paradigm and with particular knowledge and skills within the paradigm. Based on a set of scientific principles and mathematical laws, each discipline is trained to extend the laws to represent new situations.

It is not clear as to what is being taught in universities and what is being practiced in terms of theories and models of improving academic achievement. It is highly doubtful that graduate education students have been asked to solve the Glass and Smith (1978) equations or asked to replicate the results using actual statewide data. If these exercises were attempted, the flaws in the theory and mathematical model would have become apparent. The same can be said of the Hedges et al. equation. Most likely, students are never asked to test the underlying theory and model of achievement production either as a simulation or on actual data. In contrast, a fundamental part of aeronautical, mechanical, and electrical engineer training is the solving both simulated and "real" problems.

Here is a classroom exercise: The current achievement production function is:

$$A = \sum \beta D_{(Z)}$$

where A is Achievement measured in Z-scores; β is the standard regression weight; D is the explanatory variable measured in Z-scores; and Z is the Z-score. The problem: Using the information contained in these articles, sum the possible variables and estimate the value of A for $Z = 0$ and $Z = 1$. How much will achievement improve by increasing every variable by one standard deviation? What is wrong with this picture?

A three tier policymaking taxonomy was suggested in earlier articles starting with opinion, progressing to reliance on research, and ending with a comprehensive process of stating the underlying assumptions and evaluating the alternatives. The observations by Forrester tend to explain why most instructional policy-making is based on opinions (tier one) rather than on a common set of skills and knowledge developed from research (tier three). Following the thoughts of Kuhn, this is because there is not a common theory, a common set of laws, and a common methodology guiding research, which is used to prepare individuals to actually apply the theory, laws, and methodology. When there is a shortage of people with requisite knowledge and skills, opinion fills the vacuum. To use Forrester's metaphor, the crew and passengers without the requisite training are designing airplanes rather than the aeronautical, mechanical, and electrical engineers! Before this situation will change, a new set of specialized individuals must be trained. Before the new individuals can be trained, the existing examples of achievement productivity must be replaced with a more functional paradigm with a more clearly defined set of principles, knowledge, and skills.

Please return to and read the "achievement production rules." Engineers could not build aircraft under these conditions; yet schools are expected to "produce" high levels of achievement with multiple sets of ambiguous and contradictory rules. Amazingly, many schools do quite well.

A new achievement production paradigm would have similar characteristics and steps as building an airplane.

(1) What is to be accomplished—the specifications?
(2) What are the applicable laws?
(3) How is the system to be modeled?
(4) What are the initial conditions, and how should these conditions be changed?
(5) How much will the changes cost?

After repeatedly testing and evaluating various simulation models, an actual test model is carefully constructed and extensively examined. After evaluating the results and making the necessary corrections, the model is put into production. After production, the operations are continuously monitored, so improvements can be made. Increasingly, modeling is being used in many types of organizations. Is it possible for modeling to be applied in education?

Implications for Normal Science

Many of the ideas for this series of articles came from Kuhn's thoughts regarding paradigms and normal science. Importantly, these articles are not designed to reach specific conclusions regarding specific variables associated with achievement. Rather, they are designed to propose a different way of thinking about relationships—a paradigm. To follow are some relevant quotes from Kuhn with an explanation of how the proposed paradigm compares with his writing.

> By choosing "paradigm," I mean to suggest that some accepted examples of actual practice—examples which include law, theory, application, and instrumentation together—provide models from which spring particular coherent traditions of research" (p. 11).

This series of papers proposes an achievement production paradigm with an articulated theory, a mathematical law, a practical application, and instrumentation (a process of optimization). Many of the ideas spring from strengths of previous productivity research and, in some cases, apparent contradictions.

> Paradigms share two essential characteristics: 'their achievement was sufficiently unprecedented,' and 'sufficiently open-ended to leave all sorts of problems.' A paradigm 'is an object for further articulation and specification under new or more stringent conditions' (p. 23).

Clearly the theory, law, application and instrumentation is unique compared with other productivity research, and it is open-ended. There is substantial opportunity for further articulation and refinement under wide ranging conditions.

> To be accepted as a paradigm, a theory must seem better than its competitors, but it need not, and in fact never does, explain all the facts with which it can be confronted (p. 18).

Theories and mathematical models are representations of a phenomenon, and, hence, not the "real thing." Therefore, theories and models must be judged based on: (1) How well they explain the phenomenon; (2) how well they predict the outcome; and (3) how well the prediction can be verified.

A "policy behavior achievement" (PBA) paradigm is a better explanation of achievement production than a "resource achievement" prescription for a fundamental reason: Achievement is a form of behavior, and school behavior is directly influenced by policy. If, over time, the behaviors of the teacher and students change, then an improvement in achievement is likely. However, it is more likely for behaviors to change through wise policies.

Regarding the ability to predict achievement, the PBA paradigm is more accurate then the "resource achievement" prescription for several reasons. First, the PBA paradigm recognizes the ceiling effect of achievement and includes a law more accurately representing that characteristic. Second, it includes data regarding the effectiveness of existing policies even though the data are derived indirectly rather than observed. Because the effectiveness variable explains a considerable amount of the variance, its inclusion makes the predictions of achievement more accurate.

The PBA paradigm allows for, indeed requires, the testing of various theories or scenarios through the simulation process not available with other theories or models. This is possible because of the nonlinear functions enabling the use of simultaneous equations and the inclusion of cost as a variable. With a refined model identified, a comprehensive experiment can be conducted. This is not the case with existing achievement production theories and models.

> 'Normal Science' means research firmly based upon one or more past scientific achievements, achievements that some particular scientific community acknowledges for a time as supplying the foundation for its further practice. Today such achievements are recounted [by textbooks], elementary and advanced. These textbooks expound the body of accepted theory, illustrate many or all of its successful applications, and compare these applications with exemplary observations and experiments (p. 10).

Achievement production has not yet become a "normal science," as suggested by Kuhn, because there is no accepted paradigm or successful applications. Students are not asked to solve simulated problems replicating successful applications as students of engineering are asked to do.

> The study of paradigms...is what mainly prepares the student for membership in the particular community with which he will later practice (p.11).

As some point, after further articulation and refinement, the PBA paradigm could be valuable as a subject for professional training and practice.

> Men whose research is based on shared paradigms are committed to the same rules and standards of practice (pp. 10-11).

It is unclear what the current rules and standards of practice are. It is unlikely that some form of unification will take place until there is a unification of purpose among many institutions including universities, departments of education, foundations, and other organizations interested in improving the academic performance of students. For example, it is doubtful whether the various areas of education preparation—curriculum and instruction, administration, social foundations, finance—agree on common research and teaching efforts based on the same model.

> In the absence of a paradigm...all of the facts that could possibly pertain to the development of a given [phenomenon] are likely to seem equally relevant. As a result, early fact-gathering is a far more nearly random activity (p. 15).

The many contradictions in the research conclusion suggest that current fact gathering is a "nearly random activity." As the critique of the paradigm evolves, the shortcomings of the data being collected would become apparent, and there would be more specific purposes for refining the collection process.

It suggests which experiments would be worth performing (p. 18).

Based on the paradigm, there are several immediate questions worthy of further investigation:

- Is there an achievement ceiling effect?
- Is the relationship between achievement and the determinants nonlinear?
- Is there an appropriate nonlinear measurement of effect size?
- Do individual school circumstances matter in improving achievement?
- Are some schools more effective in producing achievement?
- What make these schools more effective?
- Do policies influence behavior?
- Do behaviors influence achievement?

The "Policy-Behavior-Achievement" Paradigm as Normal Science

According to Kuhn, normal science is the articulation of the theories already supplied by the paradigm. It is "the empirical work resolving some of its residual ambiguities and permitting the solution of problems to which it had previously only drawn attention" (p. 27). There are substantial questions regarding the responsibility for expanding the knowledge of the normal science of achievement production. In other disciplines, the responsibilities of the various institutions are far clearer, heavily relying on the efforts of higher education. What are the responsibilities of universities, departments of education, and other institutions interested in educational policy?

Universities are guided by three major purposes--teaching, research, service. The PBA paradigm is a possible vehicle for addressing all these purposes in preparing school policymakers. First, the necessary data for seeding the model are available from departments of education. The examples in these papers are from Minnesota Department of Education. The method to prepare the data for seeding into the model is described by the author in a 2009 article titled, "Reporting and Measuring School and District Effectiveness." The information for the profile, estimates of effectiveness, and the boundaries for the factors come from these data. Replicating this information could be a practical exercise for graduate students as a part of their statistics training, but state departments of education have the responsibility for the data and presumably for reporting this information to policymakers and the public. From my experience, there is little collaboration in this effort. Working together would be a good start.

With the necessary data available, all university departments contributing to graduate education could use the PBA paradigm to investigate the achievement policymaking process by means of the simulation model. The materials presented in the classroom, readings, and individual research would provide background for exploring various policy options. Rather than writing papers, the students would be asked to "test" the policy options using the simulation model. The very process of exploring policy options has value. The product of the exercise would be a critique of various policies, leading to the development of an achievement improvement strategy.

There are opportunities for the faculty and student to improve the paradigm by focusing on the theory, laws, applications, and instrumentation. Also, testing selected policy options in an experimental setting would also be valuable. From these experiences, a collection of case studies, valuable in the teaching process, would evolve. Even if a final testing of the strategies did not transpire, identifying and testing the underlying assumptions has value in developing skills and knowledge.

After over 25 years, my journey is at an end. It is possible to combine several seemingly unrelated aspects of achievement production into a single explanation and make predications based on that explanation. Indeed, achievement, various resources, SES, different notions of effectiveness, and cost can be coherently unified and incorporated into a method to predict changes in achievement. My original dream has been fulfilled. This is not to say that I have found THE answer, merely AN answer. It would be most gratifying if others would find better explanations and models, and better yet, use the explanations and models for training and in practice.

For those who have managed to wind their way through the morass of data and arguments, some might be disappointed because there is no definitive conclusion regarding the influence of class size or resources. Others will be disappointed because it is too complicated. Hopefully there will be a few who will see a future for these ideas. To me, the purpose was the journey and not the destination; it changed my way of thinking! Improving achievement is complex, requiring an explanation and model commensurate to the task. The ideas of the paradigm were emphasized in order to encourage researchers, trainers, and practitioners to broaden their thinking away from the traditional issues—lower class size or more money—to the holistic issue: How can a complex organization be designed and operated to reach its achievement goals? As it has been emphasized repeatedly, the focus must be on critiquing the underlying principles and not accepting "common-wisdom" conclusions.

Like Newton, "I stood on the shoulders of giants," such as Henry Levin, Herbert Walberg, Eric Hanushek, John Taylor, Thomas Kuhn, Linus Schrage, and Hilary P. Williams. Ironically, Glass and Smith were instrumental in molding my thinking (even though we disagree on the conclusions). I benefited substantially from their ideas and incorporated them freely. They deserve credit for building the foundation.

References

Achilles, C.M., B.A. Nye, J.B. Zaharias, and B.D. Fulton."The Lasting Benefits Study (LBS) in Grades 4 and 5 (1990–1991): A Legacy from Tennessee's Four-year (K–3) Class-size Study (1985–1989)." Project STAR. Paper presented at the North Carolina Association for Research in Education, Greensboro, North Carolina, January 14, 1993.

Addonizio, Michael, and James L. Phelps. "How Much do Schools and Districts Matter? A Production Function Approach to School Accountability," *Educational Considerations* 33 (Spring 2006): 51-62.

Bohrenstedt, George W., and, Brian M. Stecher, ed. *Class Size Reduction in California: Early Evaluation Findings, 1996-98*. CSR Research Consortium. Sacramento, CA: California Department of Education, June 1999. http://www.classize.org/techreport/index.htm.

_____. *Capstone Report: What We Have Learned about Class Size Reduction in California*. CSR Research Consortium. Sacramento, CA: California Department of Education, August 2002. http://www.classize.org/techreport/index-02.htm.

Forrester, Jay W. "Systems Dynamics: Future Opportunities." In *Studies in Management Science*, ed. Augusto A. Legasto, Jay W. Forrester, and James M. Lyneis, 7-21. Vol. 14, Systems Dynamics. Amsterdam, North Holland: 1980.

Glass, Gene V., and Mary Lee Smith. *Meta-analysis of Research on the Relationship of Class-size and Achievement*. San Francisco, CA: Far West Laboratory for Educational Research and Development, 1978.

Hedges, Larry V., Richard D. Laine, and Rob Greenwald. "Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes." *Educational Researcher* 23 (April 1994): 5-14.

Kuhn, Thomas S. *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press, 1970.

Levin, Henry M. "Cost-Effectiveness and Educational Policy." *Educational Evaluation and Policy Analysis* 10 (Spring 1988): 51-61.

_____. "Raising School Productivity: An X-Efficiency Approach." *Economics of Education Review* 16 (June 1997): 303-311.

Phelps, James L. "Measuring and Reporting School and District Effectiveness." *Educational Considerations* 36 (Spring 2009): 40-52.

Schrage, Linus E. *Lindo: An Optimization Modeling System*. San Francisco, CA: Scientific Press, 1991.

Silkman, Richard H., ed. *Measuring Efficiency: An Assessment of Data Envelopments Analysis*. San Francisco, CA: Jossey-Bass Inc., 1986.

Taylor, John R. *An Introduction to Error Analysis*. Mills Valley, CA: University Science Books, 1982.

Walberg, Herbert J. "Improving the Productivity of America's Schools." *Educational Leadership* 41 (May 1984): 19-27.

Williams, Hilary P. *Model Building in Mathematical Programming*. 2nd ed. New York: John Wiley, 1985.

Endnotes

[1] All subsequent references to Glass and Smith in this article refer to Gene V. Glass and Mary Lee Smith, *Meta-Analysis of Research on the Relationship of Class-Size and Achievement* (San Francisco, CA: Far West Laboratory for Educational Research and Development, 1978).

[2] All subsequent references to Walberg refer to Herbert J. Walberg, "Improving the Productivity of America's Schools," *Educational Leadership* 41 (May 1984): 19-27.

[3] All in standard scores with lines passing through the Z-score coordinates of 0, 0.

[4] All subsequent references to Kuhn in this article refer to Thomas S. Kuhn, *The Structure of Scientific Revolutions* (Chicago, IL: University of Chicago Press, 1970).

[5] Note that I have substituted "school system(s)" for "corporation(s)," and "modeling" for "system dynamics" in the quotation.

Acknowledgments

Over the years several friends and colleagues have listened to my thoughts and have given me valuable perspectives. They are listed below.

An indispensable companion during the writing phase has been a senior at Macalaster College in St. Paul, Minnesota, Michelle Neary. A superior student and an accomplished musician, she edited the papers for submission, and more importantly, made many perceptive suggestions for improving the clarity of logic and writing. I am proud to say she is our granddaughter. After graduation, Michelle plans to pursue a Ph.D. in chemistry.

Over the many years of pursuing these issues, my lovely, supportive, and helpful wife of 50 years frequently asked, "When will you be finished?" I can reply, "Now." However, I hope the issues and questions raised in this series of papers will motivate others to continue the journey.

My deepest appreciation to:

- Charles Greenleaf, former Michigan and White House official for education, and dear friend who got me started;
- William Milliken, former Governor of the State of Michigan who gave me my most professionally rewarding opportunity;
- Maris Abolins, Professor Emeritus of High Energy Physics at Michigan State University, and dear friend who opened my eyes to new ways of thinking;
- Doug Roberts, former state education official, former State Treasurer of Michigan, and dear friend who always encouraged my unorthodox efforts;
- John Porter, former Michigan State Chief School Officer, who started my thinking about educational accountability models;
- Wilbur Brookover, former Professor of Sociology at Michigan State University and dear friend who tutored me regarding the importance of community and family in improving achievement;
- Mike Addonizio, Professor of Education at Wayne State University, a colleague and frequent coauthor;
- Mal Katz, former state education official, who frequently challenged me on the relationship between laws of achievement and individual circumstances in education;
- Mike Moch, Professor of Business at Michigan State University, who tutored me regarding the role of policy and behavior in organizations.

Addendum:
Factor Analysis of Explanatory Variables in an Achievement Production Function

James L. Phelps

Combining explanatory variables into factors instead of using individual variables in an achievement production function is advocated in several of the articles in this special issue. The following is a brief overview of factor analysis explaining and illustrating the reasoning for this technique. There is a linchpin: Factor analysis is an aspect of regression analysis which is used to estimate the relationships between an outcome and the explanatory variables of a production function.

This idea originated from the desire to find a single number—an index—representing a school's socioeconomic status (SES). The process started with a large number of possible explanatory variables and was reduced down to just those variables making a significant and consistent contribution to predicted achievement. The SES index became a part of a comprehensive achievement production function. The initial goal was easily accomplished via regression for any one year; however, there was a substantial difference in the statistically significant variables and the magnitude of their weightings across years. There was no logical justification for these differences. As it turned out, small differences in the correlation matrix across years produced large differences in results. What were the reasons? Was there a workable alternative addressing these vagaries?

Factor analysis searches for combinations of variables—the factors—based on the common variance among variables in a correlation matrix. When a factor or factors have been previously conceptualized as being associated, factor analysis can confirm the assumption and provide an estimate of the strength of the factor(s). In other words, confirmatory factor analysis determines if conceptually associated variables are statistically related. If factors have not been previously conceptualized as being related, exploratory factor analysis identifies combinations of variables which are statistically related—the factors—and provides information helpful for the conceptualization effort.

While different in purpose, factor analysis and regression analysis share similarities. Regression estimates the relationships between an outcome and several explanatory variables, taking into consideration the relationships among the explanatory variables. Factor analysis, in contrast, estimates the relationships only among combinations of explanatory variables. Step-wise regression first identifies the single explanatory variable extracting the maximum variance associated with an outcome variable, removes this variance, and then identifies the next variable extracting the maximum variance, and so on until all independent variables are exhausted. In contrast, factor analysis identifies a combination of explanatory variables extracting the maximum variance, removes this variance, and then identifies the next combination of variables extracting the maximum variance, and so on. Each factor is orthogonal; that is, it is uncorrelated, with no linear relationship to the others.

Factor analysis is frequently used to explore combinations of statistically related variables by setting the number of factors to be identified at a minimal number and working upwards. After all, the better explanations are usually the simplest explanations. After the factors, their constituent variables, and their weightings have been identified, the task remains to place the results into some coherent conceptual framework. Factor analysis does not do this; indeed, factor analysis can produce incoherent results when there is substantial collinearity among all the variables. On the other hand, if there is no correlation among the explanatory variables, each variable is a factor, an easily understood but infrequent occurrence. Factor analysis is valuable for investigating student achievement where most explanatory variables are correlated.

The principle of factor analysis is illustrated mathematically by the simplest case of regression between an achievement variable (correlation subscript 1) and two explanatory variables (subscripts 2 and 3). The amount of explained variance (R^2) is calculated by the formula:

$$R^2 = r^2_{12} + r^2_{13} - 2\, r_{12}r_{13}r_{23} \,/\, 1\text{-}r^2_{23}$$

or

$$R^2 = (r^2_{12}/\ 1\text{-}r^2_{23}) + (r^2_{13}/\ 1\text{-}r^2_{23}) - (2\, r_{12}r_{13}r_{23}/\ 1\text{-}r^2_{23})$$

If the correlation between the two explanatory variables is zero (r_{23}), the third term in the numerator is zero (and the denominator becomes1); hence no common variance exists, and the explained variance is the sum of the two squared correlations. In other words, each variable is a factor. In contrast, if the correlation between the two explanatory variables is greater than zero, the common variance is subtracted from the sum of the other variances. Because of the common variance, the two explanatory variables form a factor; that is, the two explanatory variables work cooperatively rather than independently to influence the outcome. The degree to which the variables work together is measured by the common variance. In stepwise regression, the explanatory variable with the largest correlation with the outcome variable is entered first, and the common variance subtraction is applied to the next variable entered, overestimating the influence of the first and underestimating the influence of the second. This explains why small differences in the correlation matrix produce large differences in regression results across years. The ambiguous interpretations of the common variance compound as more correlated explanatory variables are added into the regression equation. Moreover, there is a point where additional variables are no longer significant, and thus eliminated from consideration in the interpretation. Given this statistical reality, there is a workable alternative. The unique variance for each variable and the common variance among all explanatory variables can be combined into a factor predicated on an underlying theory explaining how the individual variables work together to achieve an outcome.

The notion of factors is incorporated into an achievement production function when socioeconomic status (SES) is included in a

production function. Because there is no specific definition of SES, a combination of student and community characteristics is assembled as proxies to represent SES. The proxies are selected based on their conceptual logic, their statistical relationships among the variables, and their relationships with the outcome variable. In earlier papers, this notion of combining explanatory variables has also been applied to staff quantity with the variables of teachers, support teachers, teacher aides, and administrators, because these staffing roles work cooperatively to improve student achievement. Likewise, the variables of years experience, salary, age, and educational training are components of staff characteristics because these attributes combine to influence performance. Because of the substantial conceptual and statistical association of the variables within the concepts of staff quantity and staff characteristics, the use of factors seems logical. To further substantiate this position, these two conceptual factors—staff quantity and staff characteristics—are the foundation of confirmatory and exploratory factor analyses, addressing several questions. The examples are from a correlation matrix derived from the same data set described and used in the previous articles in this issue.

Are the proposed constituent explanatory variables related to the conceptual factor?

Tables 1 and 2 present the confirmatory factor analysis results for staff quantity and staff characteristics. The magnitude of association of the variables within the factor is measured in terms of factor loadings and amount of explained variance. The explained variance is calculated by dividing the squared factor loading by the number of explanatory variables. Only the relevant variables are included in the analysis. The factor analysis of staff quantity confirms the assumption that these staff roles are statistically associated. As might be expected, the contribution by teacher is highest, with administrators making little contribution to the explained variance. The factor analysis of staff characteristics confirms the assumption that these attributes are statistically associated. The contribution to the explained variance by graduate educational training (Masters Degree) is lower than other variables. Together, Tables 1 and 2 support the practice of combining explanatory variables into factors of staff quantity and staff characteristics for inclusion in an achievement production function.

Table 1
Factor Analysis of Staff Quantity

Variable	*Factor Loading*	*Squared*	*Percent*	*Variance*
Teacher	0.845	0.714	0.494	0.179
Administrator	0.099	0.010	0.007	0.002
Support	0.649	0.421	0.291	0.105
Aide	0.548	0.300	0.208	0.075
Sum		1.445		
Variance		0.361		0.361

When the constituent variables of both concepts are combined and analyzed, do they reasonably identify the two conceptual factors?

A separate exploratory factor analysis was conducted placing the constituent variables of both factors into a single analysis, restricted to two factors to determine if the analysis would identify the proposed factors. (See Table 3.) The analysis identified two factors, however, not the ones anticipated. Moreover, the resulting factors do not lead to a coherent explanation. Because of the collinearity of the variables, the staff characteristics overwhelmed the analysis, eliminating the staff quantity variables from consideration. This is an example of exploratory analysis where the factors do not lead to a coherent explanation.

Table 2
Factor Analysis of Staff Characteristics

Variable	*Factor Loading*	*Squared*	*Percent*	*Variance*
Years	0.767	0.588	0.274	0.147
Salary	0.755	0.570	0.265	0.143
Age	0.839	0.704	0.327	0.176
Masters Degree	0.537	0.288	0.134	0.072
Sum		2.151		
Variance		0.538		0.538

Table 3
Factor Analysis of Combined Explanatory Variables: Explained Variance of Contributing Variables

Variables	*Factor 1*	*Factor 2*
Staff Quantity		
Teacher	0.041	0.000
Administrator	0.014	0.0001
Support	0.001	0.006
Aide	0.032	0.002
Staff Characteristics		
Years	0.000	0.111
Salary	0.083	0.010
Age	0.002	0.110
Masters Degree	0.083	0.000
Sum	0.258	0.239

Table 4
Factor Analysis of Combined Explanatory Variables: Explained Variance

Variables	*Factor 1*	*Factor 2*	*Factor 3*
Staff Quantity			
Teacher	0.000	0.000	0.093
Administrator	0.000	0.025	0.000
Support	0.005	0.048	0.045
Aide	0.001	0.009	0.029
Staff Characteristics			
Years	0.111	0.000	0.000
Salary	0.015	0.079	0.010
Age	0.111	0.002	0.000
Masters Degree	0.001	0.056	0.027
Sum	0.244	0.220	0.205

When the constituent variables of both concepts are placed in the analysis, do they reasonably identify more than the two coherent factors?

An exploratory analysis was conducted on the same set of data allowing for three factors. (See Table 4.) Factor 1 incorporates years of service and age while the second factor incorporates support staff, salary, and masters degrees. The third combines teachers, support, and aides. Support is influential in both the second and third factor. All three factors are weaker in total variance than the ones previously identified. None of the factors reflect some higher-order concept. These results do not offer insights clearer than the analyses in Tables 1 and 2.

The first two examples confirm the statistical relationships among the component variables within the proposed staff quantity and staff characteristics factors. This occurs because the variables were preselected due to their logical association with the concept. In contrast, neatly formed factors do not emerge when all the variables, that are also correlated, are put into the analysis. Recall the three-variable regression formula: When explanatory variables are correlated, each explanatory variable cannot be a unique factor. This explains why regression results based on large numbers of correlated variables are most likely incoherent and conceptually unwise.

In these articles, the component variables are combined into regression factors and used to: (1) Report the standing of schools on the factors, rather than on individual variables; and (2) estimate the effectiveness of schools when these factors are statistically controlled. First, for each individual factor, the component variables are regressed against the achievement variable to obtain weightings, and these weightings are averaged over time.[1] The averaged weightings are then coefficients in an equation, representing the factor's relationship with the achievement variable. When the coefficients are entered into the equation for each school observation and evaluated, the results are a single number which best predicts the achievement. The result is an index combing the unique and common variance representing the standing for each school on each factor. This is done for SES, staff quantity, and staff characteristics. Now the achievement prediction equation has just three explanatory variables rather than a large number of variables.

Finally, the residuals of the yearly regression analysis are averaged to obtain an estimate of the school effectiveness. Averaging the residual is a common method in econometrics to estimate the fixed effect, i.e., the influence on achievement unique to each school. The details are included in this special issue.

In summary:

- Combining explanatory variables into factors for use in an achievement production function regression analysis is appropriate when the factor variables are conceptually and statistically related.
- Entering the individual explanatory variables separately into a production function regression analysis is appropriate only when the explanatory variables are conceptually independent and minimally correlated.
- Conversely, entering the individual explanatory variables separately into a production function regression analysis is problematic when the explanatory variables are conceptually related and substantially correlated.
- While helpful, factor analysis does not resolve all the issues inherent in regression analysis when a large number of variables are correlated. In these cases, a careful theoretical foundation is critical.

Throughout the special issue and this discussion, the purpose has been to link theory, evidence, and methodology to build a comprehensive and workable achievement production function. The underlying theory is based on what is generally accepted as being true: (1) Instructional staff work as a team to influence achievement; and (2) a combination of characteristics influence teacher behavior and performance. The evidence provided in Tables 1 and 2 supports the theory. Therefore, the logical method is to combine the variables identified conceptually and verified via factor analysis and use regression to obtain the weightings to construct an index for each factor. Finally, the indices representing the factors become the components of an achievement production function:[2]

Achievement = SES (9) + Staff Quantity (4) + Staff Characteristics (5) + Effectiveness

This comprehensive formulation brings a conceptual clarity, ease of explanation, coherence,[3] and simplicity not present when individual variables are the starting point of an achievement production function.[4]

Endnotes

[1] Because the weightings do not change over time, the best estimate of the true value is the average.

[2] The numbers in parentheses are the number of constituent variables in the factors.

[3] In an earlier effort, all the variables were entered into the equation, and it was virtually impossible to make a coherent explanation of the results because of the substantial correlation among the explanatory variables.

[4] With the variables included individually, there would be 18 mostly-correlated variables, with the dilemma of how to attribute the common variance and interpret the results.

ISSUES 1990-2011

Educational Considerations is a leading peer-reviewed journal in the field of educational leadership. Since 1990, *Educational Considerations* has featured outstanding themes and authors relating to leadership:

SPRING 1990: a theme issue devoted to public school funding.
Edited by David C. Thompson, Codirector of the UCEA Center for Education Finance at Kansas State University and Board of Editors of *Educational Considerations*.

FALL 1990: a theme issue devoted to academic success of African-American students.
Guest-edited by Robbie Steward, University of Kansas.

SPRING 1991: a theme issue devoted to school improvement.
Guest-edited by Thomas Wicks & Gerald Bailey, Kansas State University.

FALL 1991: a theme issue devoted to school choice.
Guest-edited by Julie Underwood, University of Wisconsin-Madison and member of the Editorial Advisory Board of *Educational Considerations*.

SPRING 1992: a general issue devoted to philosophers on the foundations of education.

FALL 1992: a general issue devoted to administration.

SPRING 1993: a general issue devoted to administration.

FALL 1993: a theme issue devoted to special education funding.
Guest-edited by Patricia Anthony, University of Massachusetts-Amherst and member of the Editorial Advisory Board of *Educational Considerations*

SPRING 1994: a theme issue devoted to analysis of funding education.
Guest-edited by R. Craig Wood, Codirector of the UCEA Center for Education Finance at the University of Florida and member of the Editorial Advisory Board of *Educational Considerations*.

FALL 1994: a theme issue devoted to analysis of the federal role in education funding.
Guest-edited by Deborah Verstegen, University of Virginia and member Editorial Advisory Board of *Educational Considerations*.

SPRING 1995: a theme issue devoted to topics affecting women as educational leaders.
Guest-edited by Trudy Campbell, Kansas State University.

FALL 1995: a general issue devoted to administration.

SPRING 1996: a theme issue devoted to topics of technology innovation.
Guest-edited by Gerald D. Bailey and Tweed Ross, Kansas State University.

FALL 1996: a general issue of submitted and invited manuscripts on education topics.

SPRING 1997: a theme issue devoted to foundations and philosophy of education.

FALL 1997: first issue of a companion theme set (Fall/Spring) on the state-of-the-states reports on public school funding.
Guest-edited by R. Craig Wood, University of Florida, and David C. Thompson, Kansas State University.

SPRING 1998: second issue of a companion theme set (Fall/Spring) on the state-of-the-states reports on public school funding.
Guest-edited by R. Craig Wood, University of Florida, and David C. Thompson, Kansas State University.

FALL 1998: a general issue on education-related topics.

SPRING 1999: a theme issue devoted to ESL and Culturally and Linguistically Diverse populations.
Guest edited by Kevin Murry and Socorro Herrera, Kansas State University.

FALL 1999: a theme issue devoted to technology.
Guest-edited by Tweed Ross, Kansas State University.

SPRING 2000: a general issue on education-related topics.

FALL 2000: a theme issue on 21st century topics in school funding.
Guest edited by Faith Crampton, Senior Research Associate, NEA, Washington, D.C.

SPRING 2001: a general issue on education topics.

FALL 2001: a general issue on education funding.

SPRING 2002: a general issue on education-related topics.

FALL 2002: a theme issue on critical issues in higher education finance and policy.
Guest edited by Marilyn A. Hirth, Purdue University.

SPRING 2003: a theme issue on meaningful accountability and educational reform.
Guest edited by Cynthia J. Reed, Auburn University, and Van Dempsey, West Virginia University.

ISSUES 1990-2011 continued

FALL 2003: a theme issue on issues impacting on higher education at the beginning of the 21st century.
Guest edited by Mary P. McKeown-Moak, MGT Consulting Group, Austin, Texas.

SPRING 2004: a general issue on education topics.

FALL 2004: a theme issue on issues relating to adequacy in school finance.
Guest edited by Deborah A. Verstegen, University of Virginia.

SPRING 2005: a theme issue on reform of educational leadership preparation programs.
Guest edited by Michelle D. Young, University of Missouri; Meredith Mountford, Florida Atlantic University; and Gary M. Crow, The University of Utah.

FALL 2005: a theme issue on reform of educational leadership preparation programs.
Guest edited by Teresa Northern Miller, Kansas State University.

SPRING 2006: a theme issue on reform of educational leadership preparation programs.
Guest edited by Teresa Northern Miller, Kansas State University.

FALL 2006: a theme issue on the value of exceptional ethinic minority voices.
Guest edited by Festus E. Obiakor, University of Wisconsin-Milwaukee.

SPRING 2007: a theme issue on educators with disabilities.
Guest edited by Clayton E. Keller, Metro Educational Cooperative Service Unit, Minneapolis, Minnesota, and Barbara L. Brock, Creighton University.

FALL 2007: a theme issue on multicultural adult education.
Guest edited by Jeff Zacharakis and Gabriela Díaz de Sabatés, Kansas State University, and Dianne Glass, Kansas Department of Education.

SPRING 2008: a general issue on education topics.

FALL 2008: a general issue on education topics.

SPRING 2009: a theme issue on educational leadership voices from the field.
Guest edited by Michele Acker-Hocevar, Washington State University,Teresa Northern Miller, Kansas State University, and Gary Ivory, New Mexico State University.

FALL 2009: a theme issue on leadership theory and beyond in various settings and contexts.
Guest edited by Irma O'Dell and Mary Hale Tolar, Kansas State University.

SPRING 2010: a theme issue on the administrative structure of online education.
Guest edited by Tweed W. Ross, Kansas State University.

FALL 2010: a theme issue on educational leadership challenges in the 21st century.
Guest edited by Randall S. Vesely, Indiana University-Purdue University Fort Wayne.

SPRING 2011: a theme issue on the National Council for Accreditation of Teacher Education (NCATE) Standard 4 – Diversity.
Guest edited by Jeff Zacharakis and Joelyn K. Foy, Kansas State University.