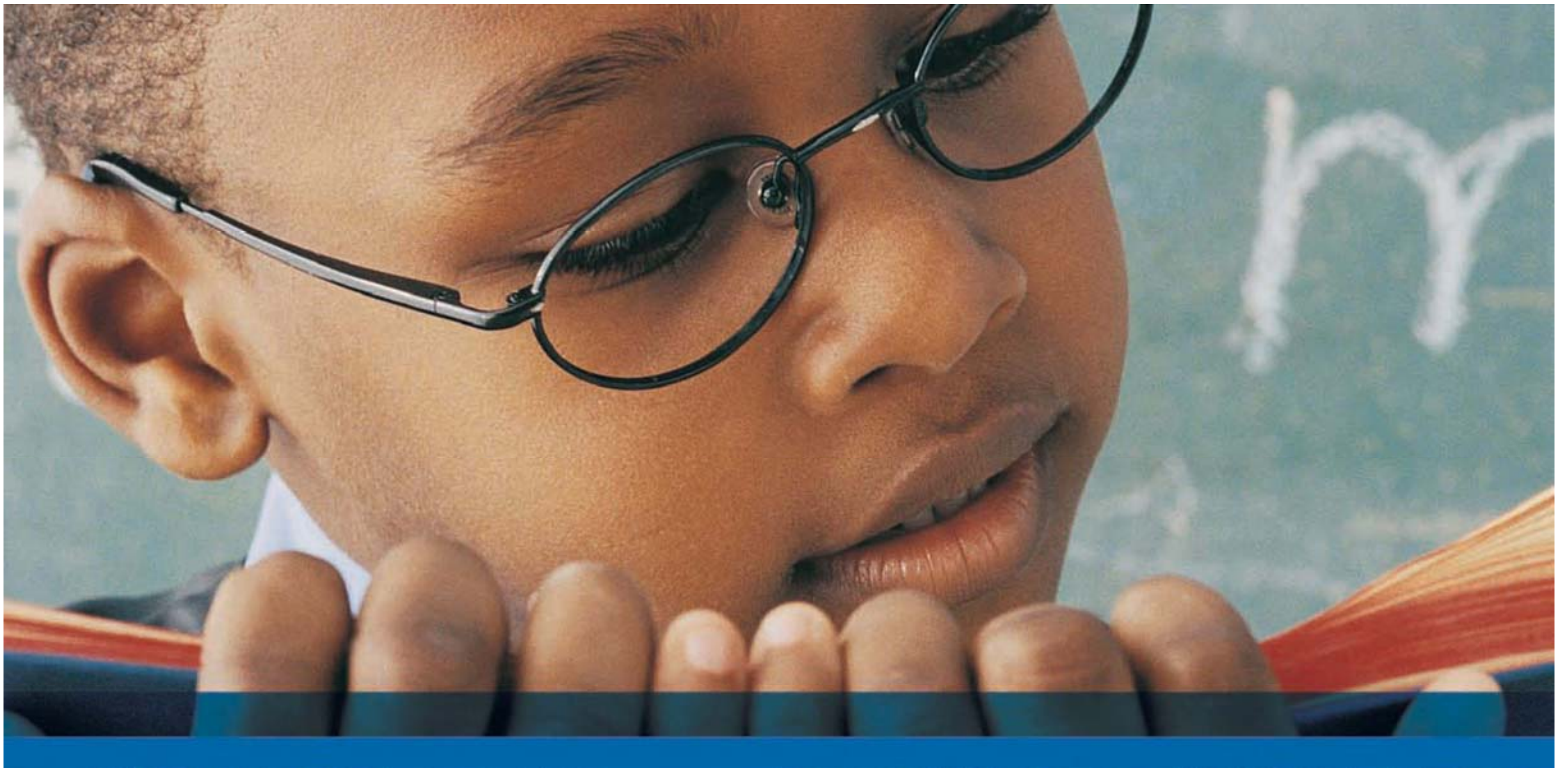

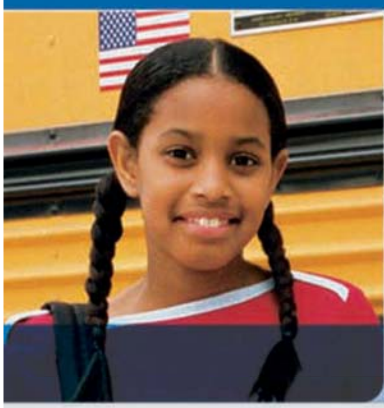

Fundamental Problems in the Measurement of Instructional Processes: Estimating Reasonable Effect Sizes and Conceptualizing What is Important to Measure

Lawrence Lanahan • Daniel J. McGrath • Mary McLaughlin • Marisa Burian-Fitzgerald • Laura Salganik

AMERICAN INSTITUTES FOR RESEARCH ®

1000 Thomas Jefferson Street, NW | Washington, D.C. 20007-3835 | 202.403.5000 | www.air.org

Table of Contents

Developing reliable and valid national-level survey measures of instructional processes and other classroom conditions related to learning has proven exceptionally difficult. Extant measures have demonstrated, at best, small links to achievement and most have not been linked statistically to achievement at all. These small links to achievement may reflect, in part, reality: instruction competes with many other factors, including factors outside of the school, for influence on student achievement. However, struggles to link to student achievement reflect also problems with the measures, including a host of validity, reliability, and generalizability problems, as well as underdeveloped conceptualization of what is important to measure about instruction and other classroom conditions. Prior research has documented validity, reliability, and generalizability problems with survey measures of instruction. In this paper, we focus on the size of the impact we can expect from instruction and teachers on student achievement and on the conceptualization of measures of instruction and other classroom conditions related to learning. We begin with a brief review of prior research describing reliability, validity, and generalizability problems with extant survey measures.

BACKGROUND: RESEARCH ON METHODOLOGICAL PROBLEMS WITH SURVEY MEASURES OF INSTRUCTION

Prior research has described many of the validity, reliability, and generalizability problems (Ball et al., 1999; Mayer, 1999a, 1999b; Mullens, 1995; Mullens, 1998; Mullens & Gayler, 1999). Lanahan, Scotchmer, & McLaughlin (2004) provided a comprehensive overview of that research and critiqued current national-level survey items on instruction in terms of the problems highlighted by the literature. They highlighted several problems prevalent in national survey measures of instruction:

Burden. Many instructional processes items are burdensome to the respondent, either because they contain too many sub-items, or because they are buried in longer surveys covering a variety of topics, in addition to instruction. High burden reduces validity because teachers become less and less likely to devote the effort necessary to answer questions accurately.

Inappropriate referent periods. Many items inquire about "typical" practice. Questions about "typical" practice suffer from validity problems, as respondents have difficulty reporting typical practice without idealizing it. Items that cover instruction over a semester or year may be asking teachers to recall so far back in time that most will not be able to report accurately or reliably. Other items that ask about recent instruction, such as in a teacher's "last lesson," are not generalizable to teachers' larger body of instruction over the academic year. Essentially, the selection of an appropriate referent period pits issues of reliability (better on general questions) against accuracy (better on specific questions) and issues of both reliability and accuracy (better over brief periods) against generalizability (better over longer periods).

Use of frequency items. Teachers are often asked to report the frequency of their use of specific instructional strategies or the frequency of the occurrence of specific classroom situations. Frequency-type items tend to have low reliability, as teachers tend to report different frequencies when re-interviewed. Moreover, frequency of the use of specific practices is difficult to link to appropriateness of the use of the practices. The appropriateness of practices is dependent on many contextual factors, including how well the teacher performs the strategy.

Focus on teachers rather than students. Although the behavior of students appears to have the stronger relationship with achievement, most survey items focus on the practices of teachers. This undermines attempts to link instructional practices to student achievement.

Fundamental issues that underlie these specific problems are the importance of *quality* of instruction to student achievement and the dual requirements of *detail* and *low burden* for reliable and valid measures of instruction. First, so much of the effectiveness of instruction depends on the quality of instruction (e.g., is lecture always a bad practice if the lecturer is very good?) and, yet, one cannot expect teachers to be accurate reporters on their own quality. Second, if researchers know, in tremendous detail, the contexts of instruction, they may be able to make good assessments of the appropriateness of instructional practice. However, the detail required is well past the ability and willingness of teachers to report accurately and reliably. Lanahan, Scotchmer, and McLaughlin (2004) concluded that the reliability, validity, and generalizability problems found in national-level survey measures undermine the legitimacy of even descriptive reporting of results collected with these measures.

The exception to the dismal overall picture of survey measurement of instruction is the measurement of instructional content (or Opportunity to Learn (OTL)). Lanahan, Scotchmer, and McLaughlin (2004) found that measures of instructional content, such as those developed by Andrew Porter and his colleagues (see Blank et al., 2001; Gamoran et al., 1997; Smithson & Porter, 1994), were more successful than measures of other areas of instruction. Teachers seem to be able to report on content with relative ease and little interpretation (Burstein et al., 1995; Leighton et al., 1994).

Ultimately, Lanahan, Scotchmer, and McLaughlin (2004) recommended a major measurement development program aimed at creating data collection tools that will pick up all of the variation in student achievement thought to be impacted by variation in classroom instructional processes. This, they believed, would entail a major conceptual and methodological reorientation and intensive in-field development activity.

To embark on a conceptual and methodological reorientation, we believe two steps must be taken. First, researchers must establish how much of an impact instructional and classroom variables can realistically be expected to have on student achievement gains. The smaller the link between these variables—relative to school and non-school variables—and student achievement, the more precise and burdensome the measurement will have to be to establish the link. Second, a solid conceptual framework of instructional processes must be established so that the measures will be comprehensive and less susceptible to previously encountered methodological hurdles.

This paper provides the groundwork for improved measurement by addressing the two problems of achievement links and inadequate conceptual frameworks encountered by previous efforts to measure instructional processes. First, we attempt to establish how much a typical student gains, in terms of standard deviations, from one year to the next, as well as how much of those gains can be attributed to classroom and instructional variables. Second, we review literatures on learning in order to draw a blueprint for a new basic conceptual framework of instruction that can help focus measures of instruction on the important elements of classrooms and provide conceptual categories for those elements. Lastly, we draw implications from our findings in these two areas for future measurement and make a set of recommendations for developing improved measures.

THE LINK BETWEEN INSTRUCTION AND ACHIEVEMENT IS SMALL BUT IMPORTANT

An important first step in developing measures of effective instructional processes is determining the size of impact expected of "effective" instruction. This is important because estimated effect sizes have implications for the sample sizes required to detect effects. Small statistical links between instructional processes and student outcomes would require large sample sizes to ensure detection. Moreover, identifying reasonable effect sizes helps in making sense of the varying claims made in research on instruction. In much educational research, two studies that would appear to present findings that oppose each other often differ sufficiently in study design that clear comparisons are difficult to make. For instance, a study advocating one instructional method implicitly at the expense of another method may not compare the two methods explicitly.[1] Other studies on the same general topic may study differing populations or use alternative outcome measures. This makes it

[1] See, for instance, Coles (2001), who notes that the National Reading Panel found very few studies that explicitly compared phonics instruction with whole language instruction.

difficult to make conclusions about the relative efficacy of different instructional methods—even if studies are reporting findings in the standardized metrics of effect sizes. Determining likely impacts of instruction on student learning helps one sort through these contrary findings. One can understand, for instance, the influence of various forms of outcome measures on effect sizes. One can also make educated guesses about studies that could be outliers rather than reliable barometers of instructional impacts.

We took several approaches to estimating reasonable effect sizes for the impact of instructional processes on student achievement. Our conclusion, that we can expect effect sizes to be small, is based on a convergence of evidence from these approaches. First, we studied student achievement growth and its relationship to instruction by measuring instructional impacts on achievement growth against two distributions of achievement. Selecting the appropriate distribution against which to measure growth is an important exercise. This is an exercise we may perform unconsciously in many realms. For example, in describing human growth, we understand that 6 inches of height is great deal of growth in one year, because we compare this growth either to our sense of the distribution of height among people or to our sense of the distribution of yearly growth among people or to our sense of both distributions. Gauging the size of academic achievement growth is more difficult for most of us. For one thing, we cannot see academic growth as readily as we can see growth in height. We examined impacts on student achievement growth against the distribution of students' yearly academic growth and against the entire distribution of achievement among students at a single grade level. For reasons we discuss below, we settled on the entire distribution of achievement among students at a single grade level as the better yardstick against which to measure impacts on student achievement growth. Second, we studied the extent to which students vary across grade levels in their yearly academic growth. Third, we reviewed literature comparing the achievement growth of high school dropouts with the growth of otherwise similar students who had stayed in school. These three approaches converged on a reasonable instructional effect size on the order of .1 standard deviation of the distribution of student achievement. We then examined literature reporting larger effect sizes. We hoped to learn why these studies found larger effect sizes than seemed reasonable and if their findings should cause us to change our estimate. We found that the larger effect sizes could be explained by regular characteristics of these studies that would not cause us to adjust our initial estimate.

MEASURING IMPACTS OF INSTRUCTION: THE DISTRIBUTION OF ACHIEVEMENT GROWTH

One approach we can take to assess the impact of instruction on achievement growth is to examine a distribution of the growth in achievement among students over a given period of time and determine the proportion of that distribution of growth that can be explained by differences in instruction. Thus, students in different instructional settings may be pre-tested at the beginning of a school year and post-tested at the end of the school year. These students may range in achievement growth from 0 to 50 points on the assessment. In other words, some students fail to gain at all in achievement, while other students gain considerably more over the school year. The distribution of students' achievement growth has a mean, a variance, and a standard deviation. One could match individual students' achievement growth to measures of their instructional setting. One could then calculate the proportion of the variance in the students' achievement growth that was associated with differences in their classroom assignments.

This method, using the distribution of achievement growth, is the method most often used to calculate the impact of instruction on achievement. The method addresses well the question "How much more can we expect student A in setting 1 to achieve in a year's instruction over student B in setting 2?"

Of course, the actual calculation of this effect size is not simple. Separating achievement growth associated with instructional setting apart from achievement growth associated with other influences, such as prior achievement and family achievement can be difficult (see Rowan, Correnti, & Miller, 2002). This difficulty is exacerbated by the small size of student achievement growth over a year's instruction.

However, recent advances in growth curve modeling have improved dramatically the ability of researchers to estimate the proportion of the distribution of student achievement growth that is attributable to classroom assignments. Using data from two cohorts of students moving from first through third grades in data drawn from *Prospects: The Congressionally Mandated Study of Educational Opportunity*, Rowan, Correnti, & Miller (2002) estimated that students' classroom assignments accounted for approximately 60 to 61 percent of the variance in student achievement growth in reading and 52 to 72 percent of the variance in student achievement growth in mathematics between first and third grades (p. 1531). He converted the percentage-of-variance-explained into d-type effect sizes.[2] He estimated yearly effect sizes of .77 to .78 standard deviations (SD) for reading growth and .72 to .85 SD for mathematics growth.

A drawback of this approach is the relativity of the measure. The effect size is pinned to the distribution of achievement growth among students. Therefore, if there is sufficient variation in achievement among students, there is a potential for a large effect size. However, if all students learn a great deal or all students learn very little—in other words, if the distribution of achievement growth is narrow—then the effect size will be small. The effect size tells us nothing about whether students learned "a lot" or "a little." It merely tells us whether students varied a great deal in their achievement growth and whether that variance in growth could be attributable to classroom settings.[3]

MEASURING IMPACTS OF INSTRUCTION: THE DISTRIBUTION OF ACHIEVEMENT

A second distribution that can be used to calculate instructional impacts is the distribution of students' overall achievement level or status. Here, students' achievement growth is compared to the overall distribution of student achievement at some base point in time. The difference in mathematical terms between this approach and the "distribution-of-student-achievement-growth" approach is the denominator. In both approaches, the numerator used in calculation of the effect size is some measure of the variance of student achievement growth that is attributable to classroom assignment. In the "distribution-of-student-achievement-growth" approach the denominator is the distribution of growth. In the "distribution-of-achievement" approach, the denominator is the distribution of students' achievement at a base point in time.

In this approach, then, we describe student achievement growth in terms of movement along the initial distribution of student achievement. Students' growth over time is measured in terms of standard deviations of the initial distribution of scores. For instance, to analyze student achievement gains between fourth and fifth grade, we would first standardize the distribution of fourth grade scores. We would then standardize the fifth grade scores and superimpose them over the original distribution. We would report gains in terms of the number of fourth grade standard deviations the fifth graders had advanced since fourth grade.

[2] D-type effect size = square root of (variance in achievement growth lying among classrooms)/ square root of (total variance in achievement growth).

[3] Of course, the variance in student achievement growth attributable to classroom setting may be exactly what researchers want in some instances, and this effect size describes that well.

Having measured achievement growth against this distribution, we can then partition the variance in student achievement growth among influences, including the influence of classrooms settings. We can calculate an effect size by comparing the standard deviation of student achievement growth attributable to classroom assignment to the standard deviation of the distribution of student achievement at the base time period. Using Rowan, Correnti, & Miller's (2002) analysis of the *Prospects* data, we calculate an effect size of .16 standard deviations for yearly growth in reading achievement and .18 standard deviations for yearly growth in mathematics achievement among third graders.[4] These effect sizes are less than half the size of the impacts measured using only the distribution of student achievement growth. The effect sizes are much smaller, because the distribution of *achievement* among first graders is much larger than the distribution of yearly *achievement growth* between first and third grades.

We find this measure useful, because of its anchoring to the distribution of all that students know. For example, the measure eliminates the possibility of large effect sizes reported in cases in which no students learned a great deal. In other words, a measure based on the distribution of student achievement growth could report a large effect size if there were a small overall distribution of student achievement growth but a large proportion of that distribution that were attributable to classroom setting; the measure based on the distribution of student achievement would not report a large effect size in this case. The measure is also useful because it is easily additive. Each year's gain can be directly added to the previous year's gain. If an effect size were .16 SD per year, then three years of favorable placements could net a student .48 standard deviations on the original distribution of student achievement.

Of course, this classroom impact includes all the influences included in classroom settings, including the influence of teachers and the influence of fellow students. The influence of fellow students could be large. One of the most enduring findings of the past thirty-five years has been the influence of peers on student achievement (Coleman et al., 1966). Thus, the effect of teachers should be smaller than the .16 – .18 SD impact of classrooms.

[4] Rowan (April 18, 2002, personal communication) provided the numerators (the standard deviations of the growth in achievement in reading and mathematics attributable to classroom assignment). We drew the denominator (standard deviation of the distributions of first grade achievement in reading and mathematics) from Phillips, Crouse, and Ralph (1998: 268).

ACHIEVEMENT GROWTH ACROSS GRADE LEVELS

Our .16 - .18 SD effect size also represents third graders only. Effects may differ across grades. We examined this possibility by studying the extent to which student achievement growth varies across grades. The major comprehensive work in this field has been performed by Dana Keller (1995). Combining fourteen estimates of academic achievement from four national assessment databases—the National Assessment of Educational Progress (NAEP), the Longitudinal Study of American Youth (LSAY), High School and Beyond (HS&B), and Project Talent (Talent)—Keller calculated the average annual gain in achievement for students between fourth and twelfth grades in mathematics, reading, science, and social studies**.** Keller determined that students learned, on average about .22 standard deviations (SD) each year (or 9 to 10 percentiles for students at the mean of the distribution) (Keller, 1995). In other words, students typically moved only a short distance along the overall distribution of the scores of their peer group.

Of course, the mean achievement growth tells us only the average growth of students. The mean achievement growth does not tell us if some students have progressed a great deal in a year and other students not at all. However, Keller found that the standard deviation of achievement growth was just .09 SD among fourth through twelfth graders. So, the distribution of achievement growth is fairly narrow. At one standard deviation beyond the mean growth, students gained just .31 SD. Converted into percentiles, this means that 84 percent of students grew .31 SD or less in a given year.

Average achievement growth varies by grade level, race/ethnicity, and prior achievement (see also Alexander, Entwistle, & Olson 2001; Entwistle, Alexander, & Olson, 1997). Keller (1995) found that students in the earlier years of school tended to make larger yearly achievement gains than did students in later grades. Blacks and Hispanics tended to score lower than Whites at the fourth grade base year. The gap between Blacks and Hispanics and Whites tended to grow between the fourth and twelfth grades (see also Ralph, Keller, & Crouse, 1994). Similarly, students who scored lower in fourth grade often fell farther behind as they progressed through twelfth grade.

COMPARING DROPOUTS AND STUDENTS

We used one additional method for estimating effect sizes for instruction. We reviewed studies that examined differences in achievement growth among students who did and did not attend school. This method does not isolate the impact of classrooms or instruction from other school-related impacts. However, it provides an estimate of the role of instruction—how much do students actually gain from attending school?

Using the High School & Beyond (HSB) study base year (tenth grade) and first follow up (twelfth grade) data, Alexander, Natriello, and Pallas (1985) compared the achievement growth of dropouts and youth on track to graduate high school on time. Controlling for prior achievement and measures of attachment to school, the authors modeled the potential achievement growth of dropouts had they remained in school and, correspondingly, on-track students had they dropped out. For both groups of students, the impact of schooling was about .10 standard deviation over the last two years of high school on a composite score HSB assessments. This represented one-third of the growth of students who had remained in school.

In summary, we estimated .16 - .18 SD as a reasonable effect size for classroom effects in the early elementary grades of schooling. This impact includes the impact of instruction as well as other important influences, such as classroom peers. During the high school years, .10 SD appears to be a reasonable estimate of the difference between attending school and dropping out of school. Of course, this impact, too, includes the effects of peers and other classroom and school influences. Finally, we learned that students grow, on average, .22 SD a year between grades 4 and 12. The standard deviation of this growth is .09 SD—indicating that 84 percent of students grow .31 SD or less per year.

Combining this evidence, we conclude that reasonable effect sizes for instruction should be on the order of .1 standard deviation and may be smaller at later grade levels. Of course this .1 standard deviation is a rough estimate of a reasonable effect size. However, given the limits on the amount of growth students exhibit per year, instruction effects are unlikely to be a great deal larger.

CLAIMS OF LARGE INSTRUCTIONAL EFFECTS

Some researchers claim to have found larger teacher effects (e.g., Jordan, Mendro, & Weerasinghe, 1997; Sanders & Rivers, 1996). However, our analysis suggests that these large effects involve, in some cases, the confounding of other classroom effects with instructional effects. In other cases, studies do not measure outcomes or select populations or treatments consistent with the outcomes, populations, and treatments of interest for national-level surveys.

Among the most notable claims of large teacher effects on student achievement are those from the research of William Sanders and associates working with a large database of Tennessee students and teachers. The Tennessee Value-Added Assessment System (TVAAS) database includes approximately three million records encompassing the entire Tennessee population of students in grades 2 through 8 and their teachers, over the period between 1990 and 1996 (Sanders & Rivers, 1996). The database links teachers to students' scores on the Tennessee Comprehensive Assessment Program (TCAP), which includes sub-tests in reading, language, mathematics, science, and social studies subject areas.

Sanders and associates have used this impressive collection of data to compare teachers in terms of the achievement growth of their students. They have not reported standardized effect sizes of the differences among teachers. However, we estimated an effect size from Sanders and Rivers' (1996) report examining the percentile gains of mathematics students placed in classrooms of varying effectiveness over a three-year period in two metropolitan school systems. The authors split teacher populations in each of grades three through five into quintiles based on the average mathematics achievement growth of students in their classrooms. They then compared the mean mathematics scores of students at the end of fifth grade for each of the 125 combinations of placements over the three years (lowest quintile—lowest quintile—lowest quintile vs. lowest quintile—lowest quintile—second quintile, etc.). They presented the results in terms of percentile gains accumulated over the three-year period by students in each of the 125 combinations.

We estimated an effect size for fifth grade teachers by comparing the mean scores of students who differed only in their placement in the last year (e.g., a three-year placement pattern of middle quintile—middle quintile—middle quintile compared with placements of middle quintile–middle quintile–highest quintile). We calculated an effect size by dividing the difference in achievement growth among students in two different placements by the difference in the placements. The difference in terms of placement between the middle and highest quintiles is the difference between the 50^{th} and 90^{th} percentiles (the differences between the midpoints of each quintile), or 1.3 SD on a normal distribution. Thus, we said that the difference in placements, the denominator for our calculation of effect size, was 1.3 SD. Next, we turned to our equation's numerator—the difference in achievement growth. The difference between mean scores of students placed two quintiles apart in fifth grade, according to Sanders and Rivers, averaged 15 percentage points on the distribution of student mathematics achievement. We converted these 15 percentage points into standard deviations. At the 50^{th} percentile of a normal distribution, 15 percentage points is .75 SD. The numerator for our equation, then, was .75 SD. Therefore, we estimated the effect size as .75 SD/1.3 SD = .58 SD. This is a large effect, suggesting placement in a setting one standard deviation better than another setting can help students move .58 standard deviations across the distribution of achievement in mathematics.

However, this effect size potentially includes a host of other influences on achievement besides the teacher. For example, the Sanders model does not include controls for differences in achievement growth among students from varied SES and racial/ethnic backgrounds or among students with different levels of prior achievement. Sanders and associates claim that these characteristics of students are unrelated to achievement growth in the TVAAS population (Sanders & Horn, 1998). However, others have suggested that the lack of correlation between prior achievement and gains actually results from a method of calculating achievement gains that forces prior achievement to be uncorrelated with gains (Glass, 1995). Our review of the literature on student achievement suggested students did vary in achievement growth by prior achievement, race/ethnicity, and SES (Alexander et al., 2001; Alexander et al., 1985; Entwstle et al., 1997; Keller, 1995; Scheerens & Bosker, 1997).

Moreover, Sanders and associates fail to address another major influence on achievement: peers. The Sanders TVAAS approach defines effective teachers in terms of the average achievement growth of students in their classrooms. Thus, all classroom influences, including characteristics of fellow students, are bundled into the designation of teacher effectiveness. But, it has been well known for many years that peer effects on student achievement can be large (Coleman et al., 1966). Thus, we suspect that much of Sanders' "teacher" effect should be shared with the collection of students making up the classroom.

In fact, we suspect that peer effects help explain a quirk in the way that TVAAS classifies teachers. In much of the TVAAS research, the authors have designated teacher effectiveness through averaging scores in individual teachers' classrooms over three years, rather than rating teachers based on their classroom effectiveness in a single year. They report that this is in order to improve the stability of the designation (Bock, Wolfe, & Fisher, 1996). However, if the teacher were really the central influence, as Sanders and associates suggest, we would expect more stability in the designation of teacher quality across years. We suspect that the instability stems in part from differences in teachers' student groups across years. As Rowan (2000, p. 4) has argued, the effectiveness of individual teachers is not consistent over time because any effect a teacher might have is confounded with the groups of students the teacher instructs. Thus, it is difficult to identify effective teachers, because so much of their effectiveness depends on their students. The TVAAS approach does not attempt to resolve this confound and has, instead, assigned the effect entirely to the teacher. Thus, we believe the Sanders and Rivers (1996) effect size is exaggerated.

Others claiming large effects tend to be studying particular elements of instruction (see, for example, studies cited in Marzano, Gaddy, & Dean, 2000). At first glance, these findings challenge our conclusion that a .05 standard deviation effect represents a large impact for an instructional practice. However, studies of specific practices are seldom designed in a way that produce effect size estimates that address the outcomes, populations, and treatments on the large scale.

An example of the large-scale would be achievement as measured on item response theory-scaled (IRT-scaled) assessments that allow comparisons along a continuum of achievement and across nationally-representative populations (e.g., NAEP). Moreover, student achievement gains would reflect sustained, long-term growth in skills and factual knowledge. In order to capture this kind of growth, growth needs to be measured over fairly long periods, such as a school year. This would allow assessments to register growth (recall that students average .22 SD of growth in a year) and to determine that the growth is robust or sustained over time.

Studies of specific instructional practices often miss these measurement criteria. First, though these studies often employ assessment tools that are sufficiently normed and tested for reliability and validity, their assessment tools are typically focused narrowly on the skills and factual knowledge of interest to the study. This narrow focus may be entirely appropriate for the immediate purposes of the studies. However, one can expect effect sizes on narrowly focused measures of outcome to be larger than effect sizes for more general measures (Ross, 1988; Stahl & Fairbanks, 1986). Researchers have found that assessment tools unique to a given study or investigator tend to inflate effect sizes, as well (Ross, 1988). Moreover, assessment tools that are close in form to tools used in instructional treatments may produce particularly high effect sizes (Ross, 1988).

Second, studies of specific practices often measure achievement growth over short periods. For instance, an instructional treatment may be designed for a single unit of instruction. Students are pre-tested and post-tested over a period of weeks or even shorter time periods. Again, the timing of assessments in these studies may be entirely appropriate for the purposes of the studies. However, these short time periods do not allow for the testing of the robustness of the growth. Not surprisingly, studies that assess retention over longer periods often find smaller effect sizes for growth (Ross, 1988).

Sustained effects in naturally occurring classroom environments are almost certain to be smaller than the impacts of the relatively short-term, high dosage treatments typical of educational studies. In the real world of classroom practice carried out over a school year, one can expect treatment effects to lessen.

Experimental designs form a special class of effect sizes that conflict with our criteria. An experimental design allows the researcher to isolate the impact of the treatment in question and to manipulate the treatment. Isolating and manipulating the treatment enhance the ability of the researcher to enlarge an effect—these are two of the benefits of experimental designs. By manipulating treatments, researchers produce samples that no longer represent the presence of the treatment in a larger, naturally occurring population. A one standard deviation difference between control and experiment conditions represents a larger difference in the treatment than would naturally occur in the teaching population. Thus, the effect size—the impact of a one standard deviation treatment—is larger than would be found in the general population of teachers and students in the U.S.

Meta-analyses form another special class of studies of specific practices. Meta-analytic techniques, of themselves, do not produce effect sizes systematically larger than would be expected in a national population. However, in practice, meta-analyses tend to exaggerate the effectiveness of practices.[5] Meta-analyses estimate effect sizes through syntheses of sets of original studies all examining the same or similar treatments. Meta-analytic tools provide powerful means for making sense of large bodies of research. Meta-analytic techniques can produce large sample sizes, and the power of large sample sizes, through the synthesis of many studies with small sample sizes and little power. However, meta-analyses are subject to other limitations of their source studies. For instance, flawed studies that produce inaccurate results can carry those flaws into the meta-analysis. Meta-analysts typically review source studies closely and include only those studies in which they are confident. However, the flaws of many studies are difficult to find. Flaws may elude reviewers of published studies. In addition, aspects of a study design that may be appropriate for the study itself may be inappropriate for a given meta-analysis.

Meta-analyses place enormous demands on the analyst to find and select appropriate studies. Analysts must pick from among stronger and weaker studies. They must consider the make up of sample populations in each study to determine studies that are appropriate to synthesize. They also must decide which studies are examining treatments that are close enough in design to be called members of the same class of treatments. And, they must decide if the statistical models employed in source studies are adequate for inclusion in the meta-analysis. Source studies that fail to include control variables important to the meta-analysis can bias meta-analytic results (Jencks, 2000).

Finally, analysts must find all related studies. This last demand is an important source of bias, because null findings are often more difficult to publish than are positive findings. Analysts need to find these unpublished null findings. Finding these studies is difficult in and of itself. However, unpublished studies of null findings are seldom completed to the polish of published findings. This

[5] See the recent Mid-continent Research for Education and Learning (McRel) report, *What works in classroom instruction* (Marzano, Gaddy, & Dean, 2000) for emphasis on effect sizes well beyond the justification of the studies themselves.

contributes to the difficulty for the analyst in determining the quality of the study and its admissibility in the meta-analysis.

Ultimately, meta-analyses tend to produce positive results (Jencks, 2000) and can produce exaggerated effect sizes. The combination of the enormous power afforded by the large sample sizes of meta-analyses and the bias toward positive findings in studies published and otherwise found for inclusion tends to produce positive results (Jencks, 2000). The bias toward positive findings in published studies and study flaws, such as omitted variables, contribute to exaggerated effect sizes.

Unfortunately, meta-analyses tend to mask the number of design flaws and analytical errors. They also make independent secondary investigations more difficult. Thus, their overall effect in education research has typically clouded confused scientific efforts to sort through a set of findings for valid conclusions.

SUMMARY: SMALL BUT POTENTIALLY IMPORTANT EFFECTS

We estimate .1 SD as a reasonable effect of instructional processes on student achievement. We determined this range by three methods. First, Rowan's (personal communication, 4/18/2002; see also Rowan, Correnti, & Miller, 2002) calculation of an effect size for the classroom assignment of third grade mathematics students in the *Prospects* sample set an initial estimate at .16 - .18 SD. We suggested this estimate represented the high end of reasonable effect sizes, because it represented young students, who may show larger effects of instruction, and because the classroom assignment impact includes influences other than teachers and instruction (such as the notable impact of fellow students). Second, we reviewed average growth of students across grade levels. Keller's (1995) review of growth in large sample, IRT-scaled assessments such as NAEP and HS&B estimated an average growth for students across grades fourth through twelfth of .22 SD of the initial fourth grade distribution of student achievement. The standard deviation of this average yearly growth was .09 SD. Growth varied somewhat by grade level (younger students tended to gain more), race/ethnicity (Blacks and Hispanics tended scored lower than Whites at fourth grade and the gap grew slightly across grades), and prior achievement (students who scored lowest in fourth grade tended to gain the least over grades). Finally, we reviewed research that estimated impacts of schooling by comparing dropouts with otherwise similar students who remained in high school. Alexander, Natriello, and Pallas (1985) estimated a .10 SD impact for schooling compared with no schooling. We concluded by estimating .1 SD as a reasonable effect size for instructional processes.

In social science, an effect size of .2 SD is considered small. It is natural to wonder, then, is it worth finding and studying measures of instructional processes related to instruction? Yes. First, among the major influences on achievement, instructional processes are the most easily manipulated by education policy. Instructional processes may be difficult to change (see, for example, Cuban, 1993), but the other major influences on achievement, such as student characteristics and peer characteristics, are much more difficult to change. Second, though the impact of instruction on achievement is small in social science terms, it is relatively large in terms of the average growth of students in a given year. The difference between effective and ineffective instruction of a school year may be equivalent to a full year's average growth. Finally, the expected effect size is not too small to detect with the sample sizes typically drawn in large studies with nationally representative samples.

LEARNING INVOLVES PRODUCTIVE ENGAGEMENT OF STUDENTS WITH APPROPRIATE CONTENT; TEACHING IS KNOWLEDGE-INTENSIVE

Despite the many divisions among researchers who study learning processes and the sometimes dubious science practiced, one can sift through the research and theory, claims and counterclaims and find foundational statements that appear to be true about learning in humans. There are two main foundations that are largely unchallenged and upon which we can begin to build a model of learning that we can use in developing measures of instruction related to learning. The two foundational statements are these:

- Learning is about the generation of knowledge within the context of prior knowledge.
- The learner is the central and most important actor in what is an active process of change.

In addition, careful analysis of competing claims and the evidence behind them yields several more statements that we can defend as part of a sensible core model of learning. For instance, by studying competing models of learning closely, one can identify some differences as unimportant to instruction. Examining the research and weighing evidence can resolve other disagreements. Differences do remain, but some of these produce interesting testable hypotheses.

In this section, we describe the bodies of literature we reviewed concerning learning and the steps we took to analyze the literature and develop a set of statements about learning. We then describe these statements and discuss their implications for instruction in schools.

We began our study of learning with a comprehensive review of major models of learning. We chose four learning models that had been used extensively over the last century and retained currency today. These models include the theories of Piaget and the associated theory of constructivism; the theories of Vygotsky and his followers; behaviorism, or operant conditioning, the learning theories most often associated with Thorndike, Watson, and Skinner; and Bandura's social learning or social cognitive theory.

Briefly, the first of the four major learning models draws on the work of Jean Piaget, a Swiss psychologist who worked through most of the twentieth century. Much of his early work was based on close observations of his own children (Piaget & Inhelder, 1969). Later researchers in the tradition have used more sophisticated techniques to test, refine, and extend the model (see, for example, Ginsburg & Opper, 1988; Iran-Nejad, 1995). The constructivist movement is associated with Piaget. Piaget's basic tenets revolve around the individual learner. The individual learns through encountering disequilibria between how she thought the world was and what new information is telling her about the world. The learner's cognitive structures engage with new information and fit the new information into her cognitive structures by *assimilating* the new information into her existing cognitive structures (or understanding of the world) or by adjusting her cognitive structures to *accommodate* the new information. In this way, the learner *constructs* new knowledge.

A second major learning model, and one that, in various forms, is the dominant model among mainstream educational researchers today, is associated with Lev Vygotsky, a Soviet developmental psychologist who worked mainly in the 1920s and early 1930s. Vygotsky conceived of learning as occurring mainly through socialization processes. The world encountered by the learner is, essentially, a social world. In addition to customs, values, and the tools needed to survive in the learner's world, everything else about the learner's world, including the physical setting, forms a socio-historically specific world. The individual learns through adapting to this social context. An important distinction from Piaget is that, in Vygotsky's model, the learner is not required to learn through repeating the discovery and disequilibria of others. In this way, Vygotsky's model is not, strictly speaking, a constructivist model. In Vygotsky's model, people learn a great deal through listening and watching; people learn through working and, in other ways, participating with others; even when working alone, people are using tools and knowledge developed by others. Vygotsky's writings have been reinterpreted by many over the years and his theories tested and built upon (Hogan & Tudge, 1999; Moll, 2001; Rogoff, 1990; Tudge & Winterhoff, 1993).

The main terms associated with his theory are *scaffolding* and *zone of proximal development*. Both refer to the same phenomenon. A learner is thought to increase his knowledge and skills through working with someone (usually an adult) more expert than himself at a level that is beyond what he knows or can do by himself but not higher than what he can do with the help of this expert

(Tudge & Winterhoff, 1993). The expert and the learner's prior knowledge, here, provide scaffolding that allow the student to complete the tasks he could not do alone. The learner and expert are working in the learner's zone of proximal development (Vygotsky, 1978). Researchers who advocate apprenticeship models of instruction are typically in the Vygotskian tradition (Lave & Wenger, 1991; Rogoff, 1990).

Behaviorism has played a large role in learning research over the past century, though it has largely fallen out of favor in recent years. In the behaviorist model of learning, learning processes are, essentially, stimulus-response processes. Learning occurs through conditioned response to stimuli (Skinner, 1963; Watson, 1917). As humans seek pleasure and avoid pain, they learn/are conditioned to the appropriate responses for given stimuli.

An advantage of this model is its inherent testability, at least for simple learning processes. To the extent that the learning environment can be controlled, stimuli can be measured and tested against learning outcomes. The process-product research prevalent in the 1970s was largely based on behaviorist principles of learning.

The behaviorist model has been most successful in describing rudimentary forms of learning that result directly from experience (Bandura, 1977; Rochat, 2001). However, even by two months of age, human learning processes tend to be too complex for the behaviorist model alone to describe adequately (Rochat, 2001). At two months, most infants begin to become reflective. This vastly complicates attempts to measure most learning stimuli experienced by the children. Through reflection, children modify initial stimuli and it becomes difficult to measure the stimuli ultimately experienced.[6] By most children's second year, their advances in symbolic thought have increased their capacities for reflection to the point that stimulus-response conditioning can play only a minor role in understanding most learning (Rochat, 2001).[7]

Finally, Albert Bandura's work has been influential in research and thought about learning over the last thirty years. With "social cognitive theory" (previously called "social learning theory"), Bandura branches out from the operant conditioning of behaviorism to an understanding of the role of social and cognitive processes and the agency of the learner in learning processes (Bandura, 1977, 1986, 2001; Grusec, 1992). In this theory, most human learning is occasioned by the maturation of cognitive capacities (e.g., capacities to attend to and retain information) and the presence of models

[6] Of course, the stimulus-response model may govern the reflections of the children. But the measurement of these internal reflections would be exceedingly difficult.

[7] Some behaviorist researchers have been able to account for cognitive processes by suggesting that these processes synthesize stimulus-response experiences over time. Thus, they can describe behavior as a response to a stimulus that has been averaged in some way (Bandura, 1977).

for behavior. Learners engage these cognitive capacities selectively, driven by motivators such as external and internal sources of reinforcement and their own sense of efficacy.

Social cognitive theory is based on three main themes. First, the human capacity for symbolic representation is an enormous learning resource. Symbolic representation provides tremendous leverage for learning. It allows people to store in memory, as images or verbal symbols, information that can be used as guides for future action (Bandura, 1977, p. 13). Verbal symbols, in particular, can help people reduce and store information for quick future retrieval. The ability to store and manipulate symbols is necessary for reflective and intentional thought. And symbolic representation allows humans to work out solutions to problems through enacting mentally, rather than physically, the outcomes of potential solutions.

Second, most human learning is vicarious. Humans do not learn most of what they know and can do through discovery. We learn mostly through observation and practice. Learning is more efficient through observation, rather than trial and error (Bandura, 1977, p. 12). Moreover, in many cases, learning through trial and error is dangerous. For example, we do not expect people to learn to drive through discovery. Instead, we model for them how to drive and we show them the consequences of good and bad driving. Similarly, some behaviors are so complex they can be learned only through modeling and practice. The acquisition of language is a prime example of a complex skill set that could not be acquired without observation.

Third, self-regulation directs learning processes from beginning to end. From the selection of stimuli to attend through the monitoring of attention and retention through decisions to retrieve knowledge or reproduce skills, self-regulation provides an executive function (Bandura, 1977, p. 13; Corno & Mandinach, 1983; Zimmerman & Martinez-Pons, 1986). An individual's self-regulatory functions, of course, develop over time and are influenced externally. People develop motivations through experience, often vicarious experience (Bandura, 1986; Schunk, 1984). An individual's sense of efficacy is also influenced by past experiences. However, the individual, through self-regulation, is centrally involved in the selection of experiences and the production of and interpretation of consequences (Bandura, 1986, 2001; Schunk, 1991). Individuals have a great deal of control over the external influences they encounter and how those influences are interpreted.

Current research tends to build on one or more of these four traditions. As mentioned above, much of the current classroom research is in the Vygotskian tradition. Exciting recent research on the development and functioning of the human brain employs constructivist theories (Iran-Nejad, 1995) or pursues cognitive theories akin to Bandura's (Pennington, Nicolich, & Rahm, 1995; Schunk, 1991; Singley & Anderson, 1989). Cognitive insights are also apparent in current "situativist" classroom research (Anderson, Greeno, Reder, & Simon, 2000; Anderson, Reder, & Simon, 1996; Cobb & Bowers, 1999; Greeno, 1997). Situativist research stresses the importance of context-specificity in understanding learning processes.

We reviewed the four main research traditions and the current research and looked first for areas of consensus. We then studied the contradictions between the main theories and among the findings of the more recent research. We sought to determine whether these contradictions were so serious they undermined the points of consensus. In no case did we find that points of consensus were undermined by points of contradiction. We also sought to resolve contradictions. We tried to determine if the contradictions were meaningful for us, given our interest in implications for instruction. We attempted to resolve other differences by weighing the evidence on either side of the dispute. In these cases, we did not count our resolution as the final word; instead we tried to phrase our resolution in terms of a testable hypothesis.

Ultimately, we generated the following set of findings or statements about learning.

1. Learning is a process of developing knowledge within the context of prior knowledge.

The learning of something new depends on how the learner relates this something new to what the learner already knows or knows how to do This means that a person *always* enters a new situation with prior knowledge. Infants are not born free of cognitive skills or information (Rochat, 2001). Infants are born with reflexes (e.g., sucking, grasping) that help them interact with the world. They also appear to be born with cognitive skills that help them guide and organize their perceptions of the world (Rochat, 2001). And, through experience in the womb they have gained information about the world. They develop new knowledge within the context of this prior knowledge.

Similarly, students enter each classroom with prior knowledge. All major theories of learning agree that this prior knowledge must be engaged by any new knowledge the teacher hopes the student will learn (Piaget & Inhelder, 1969; Vygotsky, 1978other cites from above). For example, students may enter a calculus classroom knowing nothing about calculus. But they will surely have knowledge of *something* (e.g., other mathematics content, typical classroom procedures, fellow students, etc.) in that classroom. What students learn will develop within the context of whatever that something is.

2. Learning is an active process; the learner is the central actor in the process. Learning requires the engagement of the learner's existing cognitive skills and information with content. As discussed above, learning depends on the learner's prior knowledge. But learning also depends on the executive functioning of the learner (Bandura, 1977, 2001). All along the learning process, the learner makes choices that bear on what she learns and whether she ever expresses what she has learned.

The learner's choices are, of course, influenced by factors other than the whim of the learner. Prior knowledge, again, is important because it influences what the learner can and wants to engage with (Bandura, 1977; Piaget & Inhelder, 1969). The learner's choices are also influenced by motivational factors, including expected outcomes from taking various actions and expectations of success or failure (Bandura, 1977; Brophy, 1983; Schunk, 1991; Skinner, 1963; Watson, 1917). In fact, the learner's sense of self-efficacy plays a major role in the self-regulation of learning processes (Bandura, 1977, 1993, 2001; Locke, Frederick, Lee, & Bobko, 1984).

Teachers are major influences on learner's choices. In the presentation of content and arrangement of learning environments, teachers constrain the learner's choices. Teachers can also motivate learners by being compelling models (Bandura, 1977) or through influencing the learner's expectations of outcomes for various actions (Watson, 1917; Schunk, 1982; Skinner, 1963) or through influencing the learner's sense of the likeliness of success. Furthermore, teachers can encourage or inhibit the use of specific self-regulated learning strategies (Corno and Mandinach, 1983).

3. Learning can be abetted by the help of others, especially those more knowledgeable than the learner. The major traditions differ in the role of others and the extent to which others can help learners learn. At one extreme, Piaget and his followers describe a limited role for others as presenting opportunities for disequilibria in the learner (Piaget & Inhelder, 1969; Tudge & Winterhoff, 1993). The teacher provides experiences that cause the learner to re-think his understanding of the world. The peer provides an opportunity to argue through a concept.

Most major traditions describe a more important role for others, especially adults such as teachers or mentors (Bandura, 1977; Cole, 1995; Lave & Wenger, 1991; Rogoff, 1990; Tomasello, 1999; Tudge & Winterhoff, 1993; Vygotsky, 1978). Teachers and others can serve as models for observational learning (Bandura, 1977). These others can also provide information more directly. For instance, students can build knowledge through listening to or watching others and processing the information (Rogoff, 1990). Teachers and others can also provide support information and skills (or "scaffolding" in the Vygotskian tradition) that help learners develop knowledge and practice and improve skills (Vygotsky, 1978).

We subscribe to the larger-role-for-others theory. The role of others, particularly more knowledgeable and skilled adults, as models, direct instructors, motivators, and mentors for learning is well established in the research literature (Bandura, 1977; Rogoff, 1990; Tudge & Winteroff, 1993; Tudge, Winterhoff, & Hogan, 1996).

4. Knowledge is highly context-specific; however, learners can employ knowledge developed in one setting in other settings. "Transfer," the ability to apply knowledge gained in one setting to another setting, has been the subject of considerable contention. It has been difficult for researchers to find evidence of transfer and many researchers believed that transfer did not occur or was very rare. The question of transfer is important, because school instruction is based on the premise that transfer occurs. Students are expected to apply school learning to their lives outside of school and beyond the school years. Moreover, educators often hope to reinforce learning by asking students to apply skills learned in one setting (for instance, mathematics class) in another setting (for instance, science class). The inability of researchers to find evidence of transfer has been troubling.

The growing consensus holds that something like transfer does occur (Anderson et al., 2000; Brown & Kane, 1988; Greeno, 1997). Whether "transfer" occurs or not is largely a semantic question for researchers in the field. Learners do adapt knowledge developed in one context in other contexts. The problem is that knowledge is so context-specific that it appears to transfer only under a limited set of conditions.

Perkins and Salomon (1987, 1989) describe two routes to transfer. One is the "low road," which involves low level knowledge. An example of knowledge amenable to the low road is learning to drive a car with a standard transmission. Through extensive and varied practice, driving with a manual shift and clutch can become second nature. Having learned to drive a VW with a standard transmission, a typical driver does not need to begin from scratch in learning to drive a Ford car with a standard transmission. The driver simply needs to learn any differences in the location of gears on the shift and become familiar with the feel of the new car's clutch. The complicated hand-foot coordination and the sense of when to shift transfer from one setting to the other. The presence of many shared cognitive elements across the tasks eases transfer (Singley & Anderson, 1989; Pennington et al., 1995).

The second route to transfer is the "high road," which involves the application of abstract principles across settings. Much of the transfer of concern in schools is high road transfer. Students are asked to demonstrate understanding of a concept by applying it in a novel situation. Or, students are asked to employ skills learned in one subject in situations presented in another subject. The high road requires the mindful attention of the learner (Pennington et al., 1995; Perkins & Salomon, 1987; 1989). The learner has to abstract the principle that transfers. In other words, the learner needs to identify the shared cognitive elements across settings (Brown & Kane, 1988; Gick & Holyoak, 1983; Pennington et al., 1995; Singley & Anderson, 1989).

Teachers can help students identify shared cognitive elements. For example, transfer may be enhanced if teachers give instructions on how material may be applied in other contexts or point out opportunities for transfer (Brown & Kane, 1988; Ghatala, Levin, Pressely, & Lodico, 1985; Gick & Holyoak, 1983). Also, transfer is less likely to occur when knowledge is taught in only one context (Bjork & Richardson-Klavehn, 1989). Therefore, presenting information in multiple contexts and using multiple examples may facilitate transfer (Brown & Kane, 1988; Gick & Holyoak, 1983; Ghatala, Levin, Pressely, & Lodico, 1985). Encouraging students to elaborate on material presented also enhances transfer (O'Donnell et al., 1985).

5. At the most general level, learning occurs through the cognitive engagement of the learner with the appropriate subject matter knowledge. This engagement, of course, is an active process for the learner. An important part of the process is the attention that students pay to learning activities. Learners need to attend to the subject matter they are expected to learn. Influences on attention include the motivations of a student, as well as the student's age (older students tend to be able to hold attention for longer periods), emotional state, and ability to self-regulate (Bandura, 2001; Caine & Caine, 1995; D'Arcangelo, 1998; Reardon, 1999; Sylwester 1995). In addition, attention is influenced by the relationship between learning activities or subject matter and students' prior knowledge. Of course, subject matter and learning activities contribute to attention, as well, through their influences on students motivations and emotional states (see, for instance, Bandura, 1977 and Piaget & Inhelder, 1969 on moderate novelty).

But learning is not solely determined by attention or time spent paying attention to instruction. The ability to recall information or perform on cue a skill that has been introduced to a student depends on the effectiveness of the cognitive processing that occurred to produce learning (as well as the appropriateness of the cue for recall). For information to be stored in long-term memory, it must be elaborated and encoded. Elaboration and reflection facilitate the transfer of knowledge to individual learning (Brown & Kane, 1988; Larson et al., 1985). Jensen describes elaboration as the "sorting, sifting, analyzing, testing, and deepening of learning..." (Jensen, 1998, p. 42). The processes through which humans elaborate new information are varied. Among these processes, we list five established in the research literature:

- ***Employing complex cognitive skills.*** Learners who apply higher order thinking strategies (e.g., analysis or analogy) while attempting to remember new information are more likely to retrieve the material later.

- ***Relating new material to personal experience***. Learners appear to process information particularly effectively for recall if they can relate the information to prior personal experience (Symons and Johnson, 1997).
- ***Limiting attention to subject to be learned***. Attention can easily be divided among various external stimuli in a classroom. Research has shown a simple reminder to students to pay attention and focus on what they are learning can aide learning (Fishback, 1999; Troyer & Craik, 2000).
- ***Repeating or rehearsing***. Though maligned in some research circles with an interest in pedagogy, simple repetition has been shown to increase retention (Diamond & Hopson, 1998). Repetition in this sense is not limited to exact or rote rehearsal. Of course, repetition can impede attention. However, repetition through artful rephrasing may induce the benefits of multiple encounters with new knowledge without taxing learners' patience.
- ***Experiencing new information through multiple senses***. Research suggests the more sensory inputs a content area shares, the more likely the memory will be retrieved (Schacter, 1996).

CONCLUSION

The research on learning processes and student achievement links leads us to three conclusions. First, to measure instruction in such a way as to capture an optimal amount of its link to student achievement, the measures must be based on a sound, comprehensive conceptual framework. Second, this framework must consider both the teacher and the learner, and it must be informed by research on learning and the role of the teacher in learning. Third, measurement of instruction must consider the difficulties previous survey measures have encountered with validity, reliability, and generalizability.

Any conceptual framework developed for the measurement of instruction must recognize the importance of the engagement of student learning processes with the subject matter being presented. Earlier measures focused almost exclusively on the teacher, rarely considering the learner. A framework built upon the research on learning processes would consider, for instance, the exposure of students to content that is delivered at a level that matches their prior knowledge through activities that are best designed to cognitively engage students to that particular content.

To avoid validity, reliability, and generalizability problems, we suggest that future measures focus on an assessment of relatively stable instructional characteristics, such as teacher knowledge. A focus on teacher knowledge would be the best way to bring the teacher—the person whose responsibility it is to provide the most fruitful occasions for students to productively engage with content—into the equation without encountering the problems of earlier measures: namely, methodological deficiencies and an inability to predict student achievement. Furthermore, the importance of the "knowledgeable other" found in the research on learning processes supports the measurement of teacher knowledge.

A framework based on learning processes and teacher knowledge, if properly developed and operationalized, would provide a more true-to-life portrait of instruction and learning; in other words, of teachers fostering and monitoring the successful engagement of students with content.

REFERENCES

Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2001). Schools, achievement, and inequality: A seasonal perspective. *Educational Evaluation and Policy Analysis*, *23*(2), 171–191.

Alexander, K. L., Natriello, G., & Pallas, A. M. (1985). For whom the school bell tolls: The impact of dropping out on cognitive performance. *American Sociological Review, 50*(3), 409–420.

Anderson, J. R., Greeno, J. G., Reder, L. M., & Simon, H.A. (2000). Perspectives on learning, thinking, and activity. *Educational Researcher, 29*(4), 11–13.

Anderson, J. R., Reder, L. M., & Simon, H.A. (1996). Situated learning and education. *Educational Researcher, 25*(4), 5–11.

Ash, M. J. & Sattler, H. E. (1973). *A video tape technique for assessing behavioral correlates of academic performance.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Ball, D.L., Camburn, E., Correnti, R., Phelps, G, & Wallace, R. (1999). *New tools for research on instruction and instructional policy: A web-based teacher log*. Seattle, WA: Center for Study of Teaching and Policy.

Ball, M. J., Davies, E., Duckworth, M., & Middlehurst, R. (1991). Assessing the assessments: A comparison of two clinical pragmatic profiles*. Journal of Communication Disorders, 24* (5–6), 367–379.

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Bandura, A. (1986). *Social foundations of thought and action*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Bandura, Albert (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist 28*(2): 117–148.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual Review of Psychology*, *52,*1–26.

Blank, R.K., Porter, A., & Smithson, J. (2001). *New tools for analyzing teaching, curriculum and standards in mathematics & science*. Washington, DC: Council of Chief State School Officers.

Bjork, M., & Richardson-Klavehn, A. (1989). On the puzzling relationship between environment context and human memory. In C. Izawa (Ed.), *Current issues in cognitive processes: The Tulane Flowerree Symposium on Cognition* (314–344) Hillsdale, NJ: Erlbaum Associates, Inc.

Bock, R. D., Wolfe, R., & Fisher, T. H. (1996). *A review and analysis of the University of Tennessee value-added research and assessment system.* Nashville, TN: Office of Education Accountability.

Borke, H. (1969). The communication of intent: A revised procedure for analyzing family interaction from video tapes, *Journal of Marriage & the Family 31*(3), 541–544.

Brewer, D.J., & Stasz, C. (1996). Enhancing opportunity-to-learn measures in NCES data. In G. Hoachlander, J.E. Griffith, & J.H. Ralph (Eds.), *From data to information: New directions for the National Center for Education Statistics* (96–901). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Brophy, J. (1983). Conceptualizing student motivation. *Educational Psychologist, 18*(3), 200–215.

Brown, A. L., & Kane, M. J. (1988). Preschool children can learn to transfer: Learning to learn and learning from example. *Cognitive Psychology, 20,* 493–523.

Burstein, L., McDonnell, L., Van Winkle, J., Ormseth, T. H., Mirocha, J., & Guiton, G. (1995). *Validating national curriculum indicators*. Santa Monica, CA: RAND.

Caine, R. N., & Caine, G. (1995). Reinventing schools through brain-based learning. *Educational Leadership, 52*(7), 43–47.

Cobb, P., & Bowers, J. (1999). Cognitive and situated learning perspectives in theory and practice. *Educational Researcher, 28*(2), 4–15.

Cole, M. (1995). Culture and cognitive development: From cross-cultural research to creating systems of cultural mediation. *Culture & Psychology, 1,* 25–54.

Coles, G. (2001). *Great Unmentionables: What National Reading Reports and Reading Legislation Don't Tell You*. Portsmouth, NH: Heinemann.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Government Printing Office.

Corno, L., & Mandinach, E. B. (1983). The role of cognitive engagement in classroom learning and motivation. *Educational Psychologist, 18*(2), 88–108.

Cuban, L. (1993). *How teachers taught: Constancy and change in American classrooms, 1890–1990.* (2nd ed.). New York: Teachers College Press.

Curtis, W. S. (1976). *A study of behavioral change in 50 severely multi-sensorily handicapped children through application of the video-tape recorded behavioral evaluation protocol.* Bureau of Education for the Handicapped. Washington, DC: U.S. Department of Health, Education and Welfare.

D'Arcangelo, M. (1998). The brains behind the brain. *Educational Leadership, 56*(3), 20–25.

Diamond, M., & Hopson, J. (1998). *Magic trees of the mind: How to nurture your child's intelligence, creativity, and healthy emotions from birth through adolescence.* New York: Penguin Putnam.

Entwistle, R. R., Alexander, K. L., & Olson, L. S. (1997). *Children, schools, and inequality*. Boulder, CO: Westview Press.

Ferguson, R. F. (1991). Paying for public education: New evidence on how and why money matters. *Harvard Journal on Legislation, 28*(2), 465–498.

Ferguson, R. F. (1998). Can schools narrow the black-white test score gap? In C. Jencks & M. Phillips (Eds.). *The black-white test score gap* (pp. 318–374). Washington, DC: Brookings Institution.

Ferguson, R. F., & Ladd, H. L. (1996). How and why money matters: An analysis of Alabama schools. In H. F. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education* (265–298). Washington, DC: The Brookings Institution.

Fishback, S. J. (1999). Learning and the brain. *Adult Learning, 10*(2), 18–22.

Frick, T., & Semmel, M. I. (1974). *Observational records: Observer agreement and reliabilities.* Paper presented at the annual meeting of the American Educational Research Association (Chicago, 1974).

Gamoran, A., Porter, A. C., Smithson, J., & White, P. A. (1997). Upgrading high school mathematics instruction: Improving learning opportunities for low-income, low-achieving youth. *Educational Evaluation and Policy Analysis, 19,* 325–338.

Ghatala, E. S., Levin, J. R., Presseley, M., & Lodico, M. G. (1985). Training cognitive strategy-monitoring in children. *American Educational Research Journa,l 22*(2), 199–215.

Gick, M. L. & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15,* 1–38.

Ginsburg, H. P., & Opper, S. (1988). *Piaget's theory of intellectual development* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall, Inc.

Glass, G. (1995). *Discussion of the Tennessee value added assessment system*. Available at http://glass.ed.asu.edu/gene/TVAAS/

Goldhaber, D. D., & Brewer, D. J. (1997). Evaluating the effect of teacher degree level on educational performance. In W. J. Fowler (Ed.), *Developments in school finance, 1996*, (pp. 197–210). Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Greeno, J. G. (1997). On claims that answer the wrong questions. *Educational Researcher, 26*(1), 5–17.

Grusec, J. E. (1992). Social learning theory and developmental psychology: The legacies of Robert Sears and Albert Bandura. *Developmental Psychology 28*(5), 776–786.

Hogan, D. M., & Tudge, J. R. H. (1999). Implications of Vygotksian theory for peer learning. In A.M O'Donnell & A. King (Eds.), *Cognitive perspectives on peer learning*. Matwah, NJ: Lawrence Erlbaum Associates.

Iran-Nejad, A. (1995). Constructivism as substitute for memorization in learning: Meaning is created by learner. *Education, 116*(1), 16–32.

Jencks, C. (2000). Educational research and educational policy: An historical perspective. In D. W. Grissmer & J. M. Ross (Eds.), *Analytic issues in the assessment of student achievement* (NCES 2000–050) (279–297). Washington, DC: U.S. Government Printing Office.

Jensen, E. (1998). How Julie's brain learns. *Educational Leadership, 56*(3), 41–45.

Jordan, H., Mendro, R., & Weerasinghe, D. (1997, July). *Teacher effects on longitudinal student achievement*. Paper presented at the CREATE annual meeting, Indianapolis, IN.

Kacmar, K. M., & Hochwarter, Wayne A. (1996). Rater agreement across multiple data collection media. *Journal of Social Psychology, 136*(4), 469–475.

Karni, V., Tanne, D., Rubenstien, B. S., & Askenasy, J. J. M. (1994). Dependence on REM sleep of overnight improvement of a perceptual skill. *Science, 265*, 679–682.

Kellam, S. G. (2000). Community and institutional partnerships for school violence prevention. In Kellam, S. G., Prinz, R., & Sheley, J.F. (Eds.), *Preventing school violence: Plenary papers of the 1999 conference on criminal justice research and evaluation-enhancing policy and practice through research, volume 2.* (1–19) Washington, DC: National Institute of Justice, U.S. Department of Justice, NCJ 180972.

Keller, D. (1995). *An assessment of national academic achievement growth.* Unpublished doctoral dissertation, University of Delaware.

Lanahan, L., Scotchmer, M., & McLaughlin, M. (2004). Methodological Critique Of Current NCES Survey Measures of Instructional Processes. Paper presented at the annual meeting of the American Educational Research Association: San Diego.

Larson, C. O., Dansereau, D. F., O'Donnell, A. M., Hythecker, V. I., Lambiotte, J. G., & Rocklin, T. R. (1985). Effects of metacognition and elaborative activity on cooperative learning and transfer. *Contemporary Educational Psychology, 10,* 342–348.

Locke, E. A, Frederick, E., Lee, C., & Bobko, P. (1984). Effect of self-efficacy, goals, and task strategies on task performance. *Journal of Applied Psychology, 69*(2), 241–251.

Lave, J., & Wenger, E. (1991). *Situated learning: legitimate peripheral participation.* Cambridge, England: Cambridge University Press.

Lapp, D., Flood, J., & Fisher, D. (1999). Intermediality: How the use of multiple media enhances learning. *Reading Teacher, 52*(7), 776–780.

Law, James, Barnett, Gillian, & Kot, Anne (1999). Coding parent/child interaction as a clinical outcome: A research note. *Child Language Teaching & Therapy, 15*(3), 261–275.

Leighton, M.S., Mullens, J.E., Turnball, K., Weiner, L.K., & Williams, A.S. (1994). *Measuring Instruction, Curriculum Content, and Instructional Resources: The Status of Recent Work* (NCES 95–11). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Love, J. A., Minis, D. L., Sandoval, J. H., & Yonge, G. D. (1977, April). *An instrument to describe college teaching: A reliability study*. Paper presented at the Annual meeting of the American Educational Research Association, New York, NY.

Luft, M. & Bemis, K. A. (1970). *Videotape techniques for establishing inter-rater reliability*. Paper presented at annual meeting of the American Educational Research Association, Minneapolis.

Luria, R. E. & McHugh, P. R. (1974). Reliability and clinical utility of the "Wing" Present State Examination. *Archives of General Psychiatry, 30*(6), 866–871.

Marzano, R. J., Gaddy, B. B., & Dean, C. (2000). *What works in classroom instruction*. Aurora, CO: Mid-continent research for Education and Learning.

Mayer, D. (1999a). Invited commentary: Moving toward better instructional practice data. *Education Statistics Quarterly 1*(2), 17–20.

Mayer, D. P. (1999b). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis 21*(1), 29–45.

Moll, L. (2001). Through the mediation of others: Vygotskian research on teaching. In V. Richardson (Ed.), *The handbook of research on teaching* (4th edition). Washington, DC: American Educational Research Association.

Mullens, J.E. (1995). *Classroom instructional processes: A review of existing measurement approaches and their applicability for Teacher Followup Survey* (NCES 95–15). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Mullens, J.E. (1998, October 31). *Validating teacher responses on instructional practice self-report surveys: Findings from a recent field test*. Paper presented at the Association for Public Policy Analysis and Management Research Conference, New York.

Mullens, J.E., & Gayler, K. (1999). *Measuring classroom instructional processes: Using survey and case study fieldtest results to improve item construction* (Working Paper No. 1999-08). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

National Council for the Social Studies (1994). *Expectations of excellence: Curriculum standards for social studies*. Washington, DC: Author. Available at: http://www.socialstudies.org/standards/stitle.html

National Council of Teachers of Mathematics, Inc. (2000). *Principles and standards for school mathematics*. Reston, VA: Author: Available http://standards.nctm.org/document/prepost/cover.htm.

National Reading Panel (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction.* Bethesda, MD: National Inst. of Child Health and Human Development.

National Research Council. (1996). *National science education standards.* Washington, DC: Author. Available at http://books.nap.edu/html/nses/html/index.html

Nuthall, G. (2000). The anatomy of memory in the classroom: Understanding how students acquire memory processes from classroom activities in science and social studies units. *American Educational Research Journal, 37*(1), 247–304.

O'Donnell, A. M., Danseraeu, D. F., Rocklin, T. R., Hythecker, V. I., Lambiotte, J. G., Larson, C. O., & Young, M. D. (1985). Effects of elaboration frequency on cooperative learning. *Journal of Educational Psychology, 77*(5), 572–580.

Perris, C. (1979). Interprofessional communicability and reliability of the Comprehensive Psychopathological Rating Scale (CPRS) as assessed by video-taped interviews. *Acta Psychiatrica Scandinavica, 60* (2), 144–148.

Perkins, D. N., & Salomon, G. (1987). Transfer and teaching thinking. In D. N. Perkins, J. Lochead, & J. Bishop (Eds.), *Thinking: The third international conference* (285–303). Hillsdale, NJ: Lawrence Erlbaum Associates.

Perkins, D. N., & Salomon, G. (1989). Are cognitive skills context-bound? *Educational Researcher, 18*(1), 16–25.

Pennington, N., Nicolich, R., & Rahm, J. (1995). Transfer of training between cognitive subskills: Is knowledge use specific? *Cognitive Psychology, 28,* 175–224.

Phillips, M., Crouse, J., & Ralph, J. (1998). Does the Black-White test score gap widen after children enter school? In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (229—272). Washington, DC: Brookings Institution Press.

Piaget, J., & Inhelder, B. (1969). *The psychology of the child* (trans. H. Weaver). New York: Basic Books, Inc.

Ralph, J., Keller, D., & Crouse, J. (1994). How effective are American schools? *Phi Delta Kappan, 76*(2), 144–150.

Reardon, M. (1999). The brain. Navigating the new reality: An exploration of brain-compatible learning. *Adult Learning 10* (2), 10–17.

Rochat, P. (2001). *The infant's world.* Cambridge, MA: Harvard University Press.

Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context.* Oxford: Oxford University Press.

Ross, J. A. (1988). Controlling variables: A meta-analysis of training studies. *Review of Educational Research, 58*(4), 405–437.

Rowan, B., Correnti, R., and Miller, R. J. (2002). What large-scale survey research tells us about teacher effects on student achievement: Insights from the *Prospects* Study of Elementary Schools. *Teachers College Record, 104*(8), 1525–1567.

Rowan, B. (2000). *Assessing teacher quality: Insights from school effectiveness research*. Paper prepared for U.S. Department of Education Expert Panel: Strategies for evaluating efforts to improve preservice teacher education, American Institutes for Research, Washington, DC, February 17, 2000.

Rowan, B., Schilling, S. G., Ball, D. L., & Miller, R., with Atkins-Burnett, S., Camburn, E., Harrison, D., & Phelps, G. (2001). *Measuring teachers' pedagogical content knowledge in surveys: An exploratory study*. Ann Arbor, MI: University of Michigan.

Rowan, B., Chiang, F. S., & Miller, R. J. (1997). Using research on employees' performance to study the effects of teachers on students' achievement. *Sociology of Education 70*(4), 256–284.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*(3), 247–256.

Sanders, W. L., & Rivers, J. C. (1996, November). *Cumulative and residual effects of teachers on future student academic achievement*. Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

Schacter, D. L. (1996). *Searching for memory: The brain, the mind and the past*. New York: Basic Books, Inc.

Schunk, D. H. (1982). Effects of effort attributional feedback on children's perceived self-efficacy and achievement. *Journal of Educational Psychology, 74*(4), 548–556.

Schunk, D. H. (1984) Self-efficacy perspective on achievement behavior. *Educational Psychologist 19*(1), 48–58.

Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist, 26*(3), 207–231.

Schwab, J. J. (1964). The structure of the disciplines: Meanings and significances. In G. W. Ford & L. Pugno (Eds.), *The structure of knowledge and the curriculum*. Chicago: Rand McNally.

Shulman, L.S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1–22.

Singley, M.K., & Anderson, J.R. (1989). *Transfer of cognitive skill.* Cambridge, MA: Harvard University Press.

Skinner, B. F (1963). Operant behavior. *American Psychologist, 18*(7), 503–515.

Smithson, J.L., & Porter, A.C. (1994). *Measuring classroom practice: Lessons learned from efforts to describe the enacted curriculum—the reform up close study*. Madison, WI: Consortium for Policy Research in Education.

Stahl, S. A., & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research, 56*(1), 72–110.

Stallings, J. A. & Giesin, P. A. (1977). *The study of reliability in observational data*. Revised version of a paper presented at the annual meeting of the American Educational Research Association (Chicago, April 1974).

Sylwester, R. (1995). *A celebration of neurons: An educator's guide to the brain*. Alexandria, VA: Association for Supervision & Curriculum Development.

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. *Psychological Bulletin, 121*(3), 371–394.

Tomasello, M. (1999). *The cultural origins of human cognition.* Cambridge, MA: Harvard University Press.

Troyer, A., & Craik, F. (2000). The effect of divided attention on memory for item and their context. *Canadian Journal of Experimental Psychology 54*(3), 161–171.

Tudge, J. R. H., & Winterhoff, P. A. (1993). Vygotsky, Piaget, and Bandura: Perspectives on the relations between the social world and cognitive development. *Human Development, 36*(2), 61–81.

Tudge, J. R. H., Winterhoff, P. A., & Hogan, D. M. (1996). The cognitive consequences of collaborative problem solving with and without feedback. *Child Development, 67*, 2892–2909.

Vygotsky, L.S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

Watson, J. B. (1917). An attempted formulation of the scope of behavior psychology. *Psychological Review, 24*(5), 329–352.

Wayne, A. J., & Youngs, P. (2001). *Teacher characteristics and student achievement gains: A review*. Paper presented at the annual meeting of the Association for Public Policy Analysis and Management, Washington, D.C., November 2, 2001.

Weiss, R. P. (2000). Brain-based learning. *Training & Development, 54*(7), 20–24.

Ziegler, A. (1978). The seasons of a writing workshop. *Teachers and Writers 10*(1), 36-41.

Zimmerman, B. J., & Martinez-Pons, M. (1986). Development of a structured interview for assessing student use of self-regulated learning strategies. *American Educational Research Journal, 23*(4), 614–628.