A large body of research, conducted over three decades following the Coleman report (1966), has failed to find a systematic relationship between school resources and student achievement (Hanushek, 1997). The so-called "education production function" studies, relied on readily measurable indicators of school resources (i.e., per pupil expenditures, teacher salary, library resources) but failed to account fully for key aspects of schooling processes that affect student outcomes. On the other hand, another branch of research, i.e., the so-called "effective schools" studies, found that desirable instructional practices (i.e., clear goals and high expectations, opportunity to learn, monitoring student progress) enhance student achievement (Lee, Bryk, & Smith, 1993; Purkey & Smith, 1983). These case-studies sought to identify elusive aspects of effective school context and process but failed to provide generalizable information on required resources as a sufficient base for policy making (Monk, 1992).

Needs for filling the knowledge gap concerning educational production function become more pressing as more state policymakers consider and adopt outcome-based school funding. While concerns in school finance reform shifted from equity to adequacy (Clune, 1994; Ladd, Chalk, & Hansen, 1999), it is not clear how states will accomplish this new goal. This necessitates an empirical search for what is an "adequate" level of instructional resources and practices. While several different methods have been proposed to set adequacy benchmarks for school funding and resource measures, every approach has limitations and thus the results should be interpreted carefully (Guthrie & Rothstein, 1999). One popular strategy is to determine a level of acceptable student
performance as adequate, and then to identify schools that achieve the desirable goals. When this happens the level of resources expended by such schools is deemed to be adequate.

While most previous studies on the adequacy of school funding relied on readily available measures of instructional resources such as per pupil expenditures on education, the spending measures were available only at the school district level. They failed to capture variations among schools within each district. Moreover, those previous studies were not able to examine how many resources were available in classrooms and how teachers used those resources to affect student achievement. At the same time, the studies were often conducted in particular states and the findings were hardly generalizable to the national level.

Despite the difficulty of collecting valid and reliable data on instructional resources and practices on a large scale, there were some attempts to collect the data from a nationally representative sample using survey methods. For example, the National Assessment of Educational Progress (NAEP) not only assesses students' academic achievement but also surveys their teachers about instructional resources and practices in classrooms so that the teacher survey results can be used to explain the student test scores. Previous analyses of the NAEP data shows that there are substantial interstate variations in instructional resources and practices and that the variations are somewhat related to achievement (see Barton, Coley, & Goertz, 1991; Raudenbush, Fotiu, & Cheong, 1998; Wong & Lee, 1998). Previous studies also showed that standards-based state education reforms have significant effects on classroom practices as measured by the NAEP data (Lee, 1998a; Lee, 1998b; Swanson & Stevenson, 2002).

The most serious concerns in doing research with the NAEP teacher survey data were those of errors of measurement and specification. One question is how to combine teachers' responses to multiple related questions and to construct objective measures from teachers' self-reports. A recent study showed that composite measures of instructional practices developed from survey instruments can be highly reliable, demonstrating both high test-retest reliability and a high correlation with direct observations of classroom instruction (Mayer, 1999). Another question is how to examine variations in those measures from the data collected through a multi-stage, complex sampling method, and to examine the multi-level relationships between teacher measures of resources/practices and student measures of achievement.

In light of these concerns, I have conducted a more systematic analysis of the 1992 NAEP teacher survey and student assessment data by (1) constructing objective measures of key instructional resources and practices and (2) investigating the ways in which the resources and practices affect student achievement in a multi-layered, complex school system. The study's objectives are to investigate the effectiveness of instructional resource allocation and use across the states and to explore the potentials and limitations of setting outcome-based standards of instructional resources and practices.

Analytical Framework

In recognition of the potential provided by calculators and computers for increasing children's mathematical power, recommendations for improving math education
often include more use of these tools in today's classrooms (NCTM, 1991). Instructional tools themselves, however, cannot develop a range of mathematical activities unless they are effectively used in classrooms. Improving teachers' knowledge and skills is essential in enhancing the quality of instructional services (Darling-Hammond, 1989; Shulman, 1987). Indeed, the current mathematics curriculum often fails to capitalize on the rich informal mathematics knowledge and understanding that children bring to instruction, and this type of curriculum in school mathematics often seems divorced from such familiar activities (see Resnick, 1987; Romberg & Carpenter, 1986). To help anchor mathematics concepts for students, it is important to present mathematics in "everyday" contexts and encourage students to work together in groups to solve problems. Thus, small-group work, using technologies and problem solving in the context of projects can be considered positive signs of implementation of many recent recommendations for the reform of school mathematics (see David, 1994; Weissglass, 1990; NCTM, 1991).

Building on the literature review, I developed an analytical framework to assess the effectiveness of instructional resource allocation and use. As shown in Figure 1, human and physical resources are allocated and used to deliver desirable instruction, which in turn affects school performance. If schools manage to allocate and use more resources but fail to improve teaching and learning, the allocation and use of instructional resources is hardly effective. This raises two interrelated research questions. First, what kinds of instructional resources may enhance quality instruction? Is the current level of school resources adequate enough to provide a desirable level of instruction? To explore those questions, I examined the relationship between instructional resources and practices (see arrows A and B in Figure 1). Secondly, what types of instructional practices may boost student achievement? Is the current level of instructional practices good enough to meet a desirable level of achievement? To probe those questions, I further examined the relationship between instructional practices and school performance (see arrow C in Figure 1).

![Diagram](image)

Figure 1: Analytical Framework for Assessing the Effectiveness of Instructional Resource Allocation and Use
Data and Methods

This study proceeded through two successive stages, that is, objective measurement and multilevel analysis for assessing the effectiveness of instructional resource allocation and use. Primary data sources were 1992 NAEP 8th grade mathematics student assessment and teacher survey data (11,290 teachers from 3,544 schools in 40 states). While the math achievement of 8th grade students attending public schools was assessed, information was also collected from the students' mathematics teachers about instructional materials and approaches currently used in the classroom. The first stage was to measure instructional resources and practices. The second stage was to link the measures to school performance. In the following sections, I explain the research methods employed at each stage.

Objective Measurement Method

The first stage was to create measures of instructional resources and practices from the NAEP math teacher survey data. I chose to apply the item response theory (IRT) to measure the level of key instructional resources and practices as reported by teachers. The basic idea of IRT theories and models is that from a set of observed responses to a set of items it is possible to derive measures or estimates of the underlying trait (Carroll, 1988). I chose to use the Rasch measurement model, among IRT models, which specifies only the position of an item on a difficulty scale (Wright and Masters, 1982; Wright & Stone, 1979). In the case of binary response items (yes/no), the Rasch measurement model specifies the probability of teacher n with measure \( b_n \) giving responses \( X_{ni} \) to item \( i \) with difficulty \( d_i \) as

\[
P(X_{ni}) = \frac{\exp(b_n - d_i)}{1 + \exp(b_n - d_i)}
\]

where \( X_{ni} = 0 \) when the teacher responds no and \( X_{ni} = 1 \) when the teacher responds yes to the item.

This idea can be extended to any item with ordered response alternatives such as an item with four response categories asking teachers about the frequency of an instructional activity (1=never, 2=monthly, 3=weekly, 4=daily). A teacher who chooses the third category can be considered to have chosen monthly over never (first step taken) and also weekly over monthly (second step taken), but to have failed to choose "daily" over "weekly" (third step not taken).

Then, the Rasch measure of teacher responses was estimated in a way that minimizes the difference between observed value (\( X_{ni} \)) and expected value (\( P(X_{ni}) \)). Both teacher trait (\( b_n \)) and item difficulty (\( d_i \)) were measured on the same logit scale. The logit is a log odds unit: the difference between a teacher trait measure and item difficulty was equal to the log of the teacher's probability of responding yes to the item.
Table 1: Items Used to Measure Instructional Resources and Practices from the 1992 NAEP 8th Grade Mathematics Teacher Survey

Physical Resource (PR) Items
1. How well does your school provide resources? (get all, most, some, none)
2. Student access to school-owned 4-function calculators? (yes or no)
3. Student access to school-owned scientific calculators? (yes or no)
4. Are computers available for your math class? (yes or no)

Human Resource (HR) Items
1. Training in estimation? (yes or no)
2. Training in math problem-solving? (yes or no)
3. Training in use of manipulatives? (yes or no)
4. Training in use of calculators? (yes or no)
5. Training in students' math thinking? (yes or no)
6. Training in number systems and numeration? (yes or no)
7. Training in measurement in math? (yes or no)
8. Training in geometry? (yes or no)
9. Training in probability or statistics? (yes or no)
10. Training in abstract or linear algebra? (yes or no)
11. Training in calculus? (yes or no)
12. Training in methods of middle-school math? (yes or no)

Progressive Instruction (PI) Items
1. How much emphasis on reasoning/analysis? (heavy, moderate, little/no)
2. How much emphasis on communicating math ideas? (heavy, moderate, little/no)
3. How often do students work in small groups? (daily, weekly, monthly, never)
4. How often do students use measurement/geometry? (daily, weekly, monthly, never)
5. How often do students use calculators? (daily, weekly, monthly, never)
6. How often do students use computers? (daily, weekly, monthly, never)
7. How often do students write reports/do projects? (daily, weekly, monthly, never)
8. How often do students write about problem-solving? (daily, weekly, monthly, never)
9. How often do students discuss math with others? (daily, weekly, monthly, never)
10. How often do students work real-life problems? (daily, weekly, monthly, never)
11. How often do students make up math problems? (daily, weekly, monthly, never)
12. How often assess students with written responses? (weekly, monthly, yearly, never)
13. How often assess students w/ projects/portfolios? (weekly, monthly, yearly, never)

Note: Response categories for each question are shown in parenthesis

BIGSTEPS, the Rasch measurement program, was used to construct measures from the responses of 8th grade math teachers to the 1992 NAEP teacher survey items. The item difficulty of all measures was scaled to have a mean of 50 in one-tenth of logit (i.e., the scale value of 1 logit = 10). Information on the availability of basic instructional materials and tools for students as well as teachers were obtained from the responses of 11,247 8th
grade math teachers to four items in the 1992 NAEP teacher questionnaire that intended to measure physical resources (see Table 1). Secondly, information on both pre-service and in-service teacher training in math content knowledge and pedagogical skills were obtained from the responses of 11,290 8th grade math teachers to twelve items in the 1992 NAEP teacher questionnaire that intended to measure human resources (see Table 1). Finally, the responses of 10,982 8th grade math teachers to thirteen items on current classroom activities in the 1992 NAEP teacher questionnaire were used to measure "progressive instruction" (see Table 1 above).

The applications of the Rasch model to teacher survey data on instructional resources and practices raise two major questions. First, what are the underlying traits that are measured from teacher responses? Second, what does the difficulty of survey items for each measure represent? First of all, I made a distinction between instructional resources (e.g., calculators or computers) and instructional practices (i.e., the way the calculators and computers are used by teachers). Having adequate resources may not necessarily guarantee proper uses for student learning. Further, I differentiated human resources from physical resources. I used the term human resources to denote the amount of instructional knowledge and skills that teachers acquire through training. In contrast, I used the term physical resources to denote the amount of instructional materials and tools that teachers or their students can access. Some schools may have relatively well-trained teachers but still suffer from a lack of adequate instructional materials or vice versa. Both human and physical resources were regarded as subsets of a composite school resources factor, and they would show some degree of correlation to each other.

Simple item response theory asserts that there shall be a single dimension of trait that underlies responses on a test. Unidimensionality may not be a reasonable assertion in the case of instructional resources and practices, if each resource item or practice item measures a separate trait—a trait unique or specific to itself. However, a unidimensionality assumption is reasonable, at least as a first approximation for data analysis. Indeed, calculators, computers and other instructional technologies available in math classrooms may be better conceived as a resource package rather than a discrete item. Likewise, teachers' training across content and pedagogical areas in math is likely to be interrelated. The construct validity of each measure was assessed by examining item fit. One application of the model was to determine which items measure the single underlying trait and to eliminate the items that measure it poorly or not at all, thus reducing the dimensionality of the test. The construct validity of measures was also checked by comparing the obtained difficulty order of the items with the researcher's expected order. Item difficulty was assumed to reflect the cost involved in each item, including not only financial but also human costs in acquiring or using those inputs for educational production. Instructional resources are less likely to be allocated and used when they are expensive and complex.

While physical and human resources were viewed as distinctive subsets of instructional resources, they were also likely to differ in their unit cost. For example, training a new teacher does not cost the same as purchasing a new computer. In order to compare results across two different tests, we need to make an assumption about the relationship between the tests. One common assumption is that the tests are equally
difficult on average for both samples. Since equal difficulty was unlikely to be the case for this comparison of physical vs. human resources, I checked the difficulties of all items through joint calibration and found that human resources had significantly higher level of difficulty on average than physical resources. Then I produced separate measures of human and physical resources with their item difficulties anchored on the joint calibration results in order to facilitate fair comparison of the two measures. This process, however, does not warrant the consistency of scales between the two measures for direct comparison.

**Multilevel Analysis Method**

The second stage was to detect and estimate differences in measured variables among schools within states and among states through multilevel analysis method. The data collected under NAEP was hierarchical in nature because schools were nested within states. Two-level hierarchical linear models (HLM) were used to examine the relationship between instructional resources, practices and achievement outcomes at the school and state levels simultaneously (see Bryk & Raudenbush, 1992). The use of HLM was designed to provide a more accurate estimation of sampling error resulting from the multi-stage, complex sampling in NAEP. The measurement error resulting from the multiple imputation of NAEP scores was also taken into account by averaging the parameter estimates obtained from the HLM analyses of five plausible values (see Appendix).

According to the NAEP sampling design, teachers were selected if they taught the student the subject in which the student was assessed. Therefore, the purpose of drawing teacher samples in the NAEP data was not to estimate the attributes of the teacher population but to correlate student performance with the characteristics of their teachers. Because there were very few math teachers sampled from each school, within-school variations in teacher-reported resources and practices were too small to examine. Thus, teacher-level measures of instructional resources and practices were matched to their students and aggregated to the school level for analysis. Likewise, student-level measures of math achievement were also aggregated to the school level.

Multi-level analyses of the 1992 NAEP data including 3,544 schools in 40 states were conducted to examine the relationship between instructional resources and practices first and then the relationship between instructional practices and academic achievement. A two-level HLM model was fitted using instructional resources (HR and PR) as predictors and progressive instruction (PI) as an outcome variable: all Level-1 predictors were centered at the group mean.

**Level-1 model (School Level):**

\[
(PI)_{ij} = \beta 0j + \beta 1j (HR)_{ij} + \beta 2j (PR)_{ij} + e_{ij}
\]

\((PI)_{ij}\) is the average measure of progressive instruction in school \(i\) in state \(j\);  
\((HR)_{ij}\) is the average measure of human resources in school \(i\) in state \(j\);  
\((PR)_{ij}\) is the average measure of physical resources in school \(i\) in state \(j\);  
\(e_{ij}\) is a Level-1 random effect that represents the deviation of school \(ij\)'s score from the predicted score based on school-level model.
Level-2 model (State Level):

\[ \beta_0j = \gamma_{00} + r_{0j} \]
\[ \beta_1j = \gamma_{10} + r_{1j} \]
\[ \beta_2j = \gamma_{20} + r_{2j} \]

\( \beta_0j \) represents the state j's average measure of progressive instruction.
\( \beta_1j \) represents the effect of human resources on progressive instruction in state j.
\( \beta_2j \) represents the effect of physical resources on progressive instruction in state j.

Subsequently, another two-level HLM model was fitted with progressive instruction (PI) as a predictor and school performance (SP) as an outcome variable: the Level-1 predictor was group-mean centered.

Level-1 model (School Level):

\[ (SP)_{ij} = \beta_0j + \beta_1j (PI)_{ij} + e_{ij} \]

\( (SP)_{ij} \) is the average measure of math achievement in school i in state j;
\( (PI)_{ij} \) is the average measure of progressive instruction in school i in state j;
\( e_{ij} \) is a Level-1 random effect that represents the deviation of school ij's score from the predicted score based on school-level model.

Level-2 model (State Level):

\[ \beta_0j = \gamma_{00} + r_{0j} \]
\[ \beta_1j = \gamma_{10} + r_{1j} \]

\( \beta_0j \) represents the state average level of math achievement.
\( \beta_1j \) represents the effect of instructional practices on math achievement.

The series of multi-level analyses provided information on the instructional resources-practices relationship as well as the instructional practices-math achievement relationship within 40 states. Each state was assumed to have its own regression equation with an intercept and a slope. Specifically, the random-coefficient models were designed to answer the following questions. First, what is the average of the 40 regression equations, that is, an overall pattern of outcomes and relationships across states? The HLM analyses provided the average of all states' intercepts and slopes along with reliability estimates. Second, how much do the regression equations vary from state to state? The HLM analyses provided estimates of the variances of the random effects and tests of the hypothesis that these variances are null. Third, what is the correlation between the intercepts and slopes? The HLM analyses provided an estimate of the correlation to tell whether states with large intercepts (e.g., high mean achievement) also have large or small slopes (strong or weak relationships between instructional practices and achievement).
Results

Measures of Instructional Resources and Practices

In Figure 2, the measures of instructional resources were laid out vertically with the highest rated teachers and the most difficult items at the top. Two different sets of resource measures were obtained for each teacher by first jointly calibrating item difficulties with human and physical resource items together and then separately estimating two teacher measures with item difficulties anchored on the combined calibrations.

```
<table>
<thead>
<tr>
<th>Teachers</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>.</td>
</tr>
<tr>
<td>Q 90</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>.</td>
</tr>
<tr>
<td>80</td>
<td>. #</td>
</tr>
<tr>
<td></td>
<td>.</td>
</tr>
<tr>
<td>S 70</td>
<td>##</td>
</tr>
<tr>
<td></td>
<td>#</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>####+</td>
</tr>
<tr>
<td></td>
<td>#</td>
</tr>
<tr>
<td>M 50</td>
<td>####+</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>####+</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>S 30</td>
<td>####+</td>
</tr>
<tr>
<td></td>
<td>##</td>
</tr>
<tr>
<td>20</td>
<td>#</td>
</tr>
<tr>
<td></td>
<td>.</td>
</tr>
<tr>
<td>Q 10</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>.</td>
</tr>
<tr>
<td>0</td>
<td>.</td>
</tr>
<tr>
<td></td>
<td>.</td>
</tr>
<tr>
<td>-10</td>
<td>.</td>
</tr>
</tbody>
</table>
```

Figure 2: Distributions of Teachers and Instructional Resource Items on a Common Logit Scale
The overall reliability coefficient of instructional resource measures was .74. The Rasch measurement provided misfit statistics as indicators of how well test data fit the model. For human resources items, the mean squares misfit statistics ranged from .88 to 1.20 ($M = 1.00, SD = .12$). For physical resources items, the mean squares misfit statistics ranged from .87 to 1.13 ($M = 1.05, SD = .10$). Therefore, all misfit statistics were within acceptable ranges.

Table 2 shows the results of calibrating progressive instruction items through the Rasch measurement method. The reliability coefficient of this instructional practice measure was .77. In addition, the robustness of misfit and point-biserial statistics across the types of items validate the construct. Items were hierarchically ordered in terms of item difficulty to define a construct of progressive instruction. Goal-related items were less difficult than evaluation-related items, whereas the difficulty of practice-related items was dispersed according to the characteristics of the activity.

Table 2: Summary Statistics of Rasch Measurement: Progressive Instruction (PI) Items

<table>
<thead>
<tr>
<th>Item No.</th>
<th>Goal (Emphasis)</th>
<th>Practice (Activity)</th>
<th>Evaluation (Assessment)</th>
<th>Measure (Error)</th>
<th>Misfit a</th>
<th>Point-Biserial b</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Goal</td>
<td>Write reports/Do projects</td>
<td>Projects/Portfolios</td>
<td>71.73 (.22)</td>
<td>.75</td>
<td>.39</td>
</tr>
<tr>
<td>6</td>
<td>Practice</td>
<td>Use Computers</td>
<td>Projects/Portfolios</td>
<td>63.40 (.16)</td>
<td>1.49</td>
<td>.13</td>
</tr>
<tr>
<td>11</td>
<td>Evaluation</td>
<td>Make up math Problems</td>
<td>Written responses</td>
<td>61.53 (.15)</td>
<td>.96</td>
<td>.41</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Use measurement</td>
<td>Written responses</td>
<td>59.19 (.14)</td>
<td>.72</td>
<td>.38</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>Projects/Portfolios</td>
<td>Written responses</td>
<td>57.50 (.14)</td>
<td>.94</td>
<td>.44</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>Write about problem-solving</td>
<td>Written responses</td>
<td>56.12 (.13)</td>
<td>.85</td>
<td>.50</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>Written responses</td>
<td>Written responses</td>
<td>50.07 (.12)</td>
<td>1.11</td>
<td>.43</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Work in small groups</td>
<td>Written responses</td>
<td>44.63 (.12)</td>
<td>.87</td>
<td>.45</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>Use calculators</td>
<td>Written responses</td>
<td>42.79 (.12)</td>
<td>1.60</td>
<td>.25</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>Work real-life math problems</td>
<td>Written responses</td>
<td>38.85 (.12)</td>
<td>.74</td>
<td>.44</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Communicating math ideas</td>
<td>Written responses</td>
<td>36.25 (.16)</td>
<td>.93</td>
<td>.41</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>Discuss math with others</td>
<td>Written responses</td>
<td>34.21 (.13)</td>
<td>.99</td>
<td>.42</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>Reasoning/Analysis</td>
<td>Written responses</td>
<td>33.73 (.17)</td>
<td>.96</td>
<td>.37</td>
</tr>
</tbody>
</table>

Note: Items are arranged and shown in difficulty order.

a Values substantially greater or less than 1 indicate that items poorly define the construct.

b The coefficient indicates a correlation between the teachers' responses to an item and their total scores (i.e., progressive instruction measure).

The difficulty of teachers' having students engage in a particular classroom activity seems to reflect the cost and complexity of implementing the activity: the more an activity requires expenses and efforts on the part of schools or teachers, the less likely teachers
are to practice it. For example, having students write reports or do projects turned out to be the most difficult-to-practice. This can be explained by the fact that the activity incurs high opportunity cost by taking up most of the class time and thus reducing expected content coverage. Using computers appears to be more difficult than using calculators because the former requires higher costs for purchase and greater complexity for operation.

Effectiveness of Instructional Resource Allocation and Use

Figures 3 and 4 show the plots of average state progressive instruction measures against instructional resources measures among the 40 states. Both figures show positive relations between the availability of key instructional resources and the level of progressive instructional practices.
The correlation between progressive instruction and teacher training \( (r = .53) \) was greater than the correlation between progressive instruction and classroom resources \( (r = .38) \). However, there was no significant relationship between the two different types of instructional resources \( (r = .12) \). This indicates that the availability of instructional tools is independent from the level of teachers' training in content and pedagogical knowledge at the state level. Thus the allocation and use of the two different types of resources for the provision of instructional practices may not be well aligned with each other.

By and large, classroom adoption of progressive instruction is fostered by the availability of such key resources as instructional tools and teachers' subject-matter knowledge. At the state aggregate level, upgrading instructional tools and improving teachers' subject-matter knowledge enhances the level of progressive instruction.

Using this progressive instruction variable, I classified states into three groups according to the teachers' and students' engagement in progressive classroom activities:
least progressive instruction = bottom quartile of states \((N = 10)\), moderate progressive instruction = middle half of states \((N = 20)\), most progressive instruction = top quartile of states \((N = 10)\). At the state aggregate level, average mathematics achievement tended to vary according to the extent to which teachers adopted such "progressive" instructional practices as student-centered, higher-order learning activities (see Figure 5).

Table 3 summarizes the results of the HLM analysis on the relationship between instructional resources and practices. The effect of human resources on progressive instruction (HR effect) was .135, whereas the effect of physical resources on progressive instruction (PR effect) was .089. All mean effects included adjustment for the other variable in the model, and all were statistically significant at probability levels less than .001. Further, the difference in effect size between these two types of resources (i.e., .135 -.089 = .064) was also statistically significant (reject H0: .135 = .089 with chi-square statistic of 5.107, \(df = 1\), \(p < .05\)). In other words, human resources were generally more cost-effective than physical resources in producing progressive instruction in math. This indicates that the current school delivery of progressive instruction was labor-intensive.

The correlations among the random effects indicate the general structure of instructional resource allocation. A high level of progressive instruction was associated with a smaller HR effect \((r = -.153)\) and a greater PR effect \((r = .555)\). This indicates that states producing more progressive instruction tend to use physical resources more effectively than human resources (i.e., PR-intensive or HR-saving). There was also a substantial negative correlation between HR effect and PR effect \((r = -.546)\). This indicates that states using physical resources more effectively tend to use human resources less effectively.
Table 3: Results of HLM Analyses: Effects of Human Resources (HR) and Physical Resources (PR) on Progressive Instruction (PI)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimated Coefficients Mean PI</th>
<th>Estimated Standard Error Mean PI</th>
<th>t-Statistic Mean PI</th>
<th>p-Value Mean PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>44.655</td>
<td>.269</td>
<td>166.17</td>
<td>.000</td>
</tr>
<tr>
<td>(Mean PI)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human Resources (HR)</td>
<td>.135</td>
<td>.014</td>
<td>9.583</td>
<td>.000</td>
</tr>
<tr>
<td>Effect</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical Resources (PR)</td>
<td>.089</td>
<td>.010</td>
<td>8.574</td>
<td>.000</td>
</tr>
<tr>
<td>Effect</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The Chi-Square Table

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimated Variance</th>
<th>Degrees of Freedom</th>
<th>Chi-Square</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean PI</td>
<td>2.184</td>
<td>39</td>
<td>277.77</td>
<td>.000</td>
</tr>
<tr>
<td>HR Effect</td>
<td>.005</td>
<td>39</td>
<td>144.23</td>
<td>.000</td>
</tr>
<tr>
<td>PR Effect</td>
<td>.002</td>
<td>39</td>
<td>66.94</td>
<td>.004</td>
</tr>
</tbody>
</table>

Correlations among Random Effects

<table>
<thead>
<tr>
<th></th>
<th>Mean PI</th>
<th>HR Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR Effect</td>
<td>-.153</td>
<td></td>
</tr>
<tr>
<td>PR Effect</td>
<td>.555</td>
<td>-.546</td>
</tr>
</tbody>
</table>

Reliability of Random Effects

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean PI</td>
<td>= .736</td>
<td></td>
</tr>
<tr>
<td>HR Effect</td>
<td>= .634</td>
<td></td>
</tr>
<tr>
<td>PR Effect</td>
<td>= .386</td>
<td></td>
</tr>
</tbody>
</table>

Table 4 summarizes the results of HLM analyses on the relationship between progressive instruction (input) and school performance (outcome). School average measure of instructional practices was significantly positively related to school average math achievement score. This indicates that an effective use of instructional resources involves more frequent student-centered, higher-order learning activities with use of modern technologies, and thus leads to an improvement of school performance.

The effect of progressive instruction on school performance turned out to vary significantly among the states. In other words, some states were better able to link instructional practices to school effectiveness. However, higher performing states did not show a stronger relationship between progressive instruction and school performance (i.e., more effective resource use); correlation between the random effects was positive but very low ($r = .268$).
Table 4: Results of HLM Analyses: Effect of Progressive Instruction (PI) on School Performance (SP)

<table>
<thead>
<tr>
<th>Estimated Effects</th>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t-Statistic</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept (Mean SP)</td>
<td>266.671</td>
<td>1.621</td>
<td>165.554</td>
<td>.000</td>
</tr>
<tr>
<td>Progressive Instruction (PI Effect)</td>
<td>.260</td>
<td>.064</td>
<td>4.037</td>
<td>.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimated Variance</th>
<th>Degrees of Freedom</th>
<th>Chi-Square</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean SP</td>
<td>99.682</td>
<td>39</td>
<td>1129.78</td>
<td>.000</td>
</tr>
<tr>
<td>PI Effect</td>
<td>.081</td>
<td>39</td>
<td>89.10</td>
<td>.000</td>
</tr>
</tbody>
</table>

Correlation between Random Effects
- Mean SP: .268
- PI Effect: .268

Reliability of Random Effects
- Mean SP = .949
- PI Effect = .485

Setting Outcome-Based Standards for Instructional Resources and Practices

The findings of this study have implications for determining adequate levels of instructional practices and resources in mathematics. The reauthorized Elementary and Secondary School Act (ESEA) in the United States, that is, the No Child Left Behind Act of 2001 (NCLB), requires standards-based accountability for public schools. One major component of this accountability policy is to evaluate whether the schools in each state are making adequate yearly progress (AYP) towards the goal of having 100% of its students reach proficient level of achievement. As the current law insists on using proficient as its target achievement level, the appropriateness of current NAEP achievement levels as a benchmark for states goes under scrutiny. The previous evaluations raised some concerns about the reliability and validity of the NAEP achievement levels and the possibility of misleading interpretations (see Shepard, Glaser, Linn, & Bohrnstedt, 1993; U.S. General Accounting Office, 1993). Moreover, there is significant variability in the definitions of proficient academic achievement by states for purposes of the NCLB (see Linn, 2003).

On the other hand, recently there were attempts to estimate the costs of raising all children's test scores up to a particular state's standard (see for reviews Guthrie &
Rothstein, 1999; Mathis, 2003). Although there were variations in the cost estimates among different studies using different methods with data from different states, they all revealed massive new investments in education spending; increases in base cost were in the 15% and 46% range (Mathis, 2003). While these estimates gave specific dollar amounts as minimally adequate level of overall school funding at the district level, they were not able to pinpoint the areas of funding where school-level or classroom-level support is needed. It is unknown how to allocate any increased fund to different types of instructional resources and how to use such newly allocated resources to improve teaching and learning.

Despite the limitations of survey method and measures used, this study's findings may provide some insight into teachers' perceived needs for instructional resources and support relative to expected student outcomes. The process of identifying desirable levels of resources and practices for classroom teachers involves tracking the input-process-outcome linkages, from student outcomes to instructional practices, and then from instructional practices to resources. Once we identify a desirable minimum level of school performance (e.g., 50% proficiency) for a given year, we can project the cut score of chosen performance level onto the scale of school input variables according to their relationships, and to pinpoint the level of school input deemed adequate.

While determining a level of achievement outcome that is considered adequate for each state is a public policy decision, schools in some states might produce any given level of output more effectively than their counterparts in other states. Figure 3 illustrates a hypothetical relationship between school input and outcome variables in two states, A and B. State A does not only produce a higher level of outcome than State B (greater intercept) but also uses fewer inputs to produce the same level of outcome (greater slope). Thus, resource allocation and use by schools in State A may be regarded as more cost-effective than State B. Consequently, the level of adequate school input needed to produce desirable performance level $Y^*$ varies between the two states: $X_a$ in State A vs. $X_b$ in State B.

![Figure 6: Hypothetical Relationships between School Input and Outcome Variables in States A and B](image-url)
For example, the following equation is derived from the HLM analysis results on the relationship between progressive instruction (X) and math achievement (Y) across states in the previous section (see Table 4): \( Y = 266.7 + .26 (X - 44.5) \) where 266.7 is the intercept, .26 is the slope coefficient, and 44.5 is the state mean of X. Suppose that we want to identify the level of progressive instruction that corresponds to the "Basic" achievement level in eighth grade mathematics as defined by the National Assessment Governing Board. The Basic level was set at a score of 262 on a 0 to 500 scale, and eighth-grade students performing at this level should exhibit evidence of conceptual and procedural understanding in math (Mullis et al., 1993). Then, the school average measure of progressive instruction as corresponding to the Basic achievement level would be 26.4. This level of progressive instruction can be interpreted in probabilistic terms based on the logit differences between 26.4 and the item difficulties as follows (see Logit-to-Probability Conversion Table in Linacre & Wright, 1997):

- The probability of having students write reports or do projects daily or weekly is only 1% (gap = -45.3).
- The probability of assessing students with projects/portfolios weekly or monthly is about 5% (gap = -31.1).
- The probability of having students work real-life math problems daily or weekly is about 25% (gap = -12.45).
- The probability of giving a heavy emphasis on reasoning/analysis is about 30% (gap = -7.3)

The above descriptions of selected item responses indicate that instructional practices corresponding to the NAEP Basic achievement level are hardly progressive. Schools whose average student performs at the Basic level would take more traditional teacher-centered approaches to instruction with emphasis on basic skills in math classes. The chance is very low that students in these schools would have opportunities to get involved in student-centered learning activities with emphasis on higher-order thinking skills.

Likewise, we can identify the level of progressive instruction that matches the "Proficient" level of 8th grade mathematics achievement. The Proficient cut score is 299, and students performing at this level should apply math concepts and procedures consistently to complex problems in math. The measure of progressive instruction that matches the achievement score of 299 is 168.7, which indicates that teachers regularly practice all of the desirable classroom activities with 100% certainty. Such extraordinary level of progressive instruction, far beyond the distribution of sample schools, may be required for schools to perform at the Proficient level on average. Nevertheless, it is difficult to extrapolate the regression line to identify the value of X associated with mean Y, because the predictive ability of the regression line falls markedly as X departs progressively from the mean of X. Thus, it is more realistic to set the performance standard closer to the average level of current student achievement in a state so that a fixed predictor value corresponding to the conditional mean outcome value can be identified with greater accuracy.
Once we have set standards for instructional practices, the same procedures can be applied to setting standards for instructional resources. Here another question is raised as to setting standards for multiple inputs that are simultaneously linked to one common outcome. Suppose we run a multiple regression of Y on several Xs and identify the unique (partial) effect of each X on Y. If the input variables were measured on a common scale with an adjustment for their probable cost differences, unstandardized regression coefficients for the inputs could be used as the indicators of cost-effectiveness. Then, the coefficient becomes a weight for each X in determining the level of each X required for producing a certain level of Y. The more cost-effective X is, the more it should be used.

To illustrate this idea, the following equation is derived from the HLM results on the relationships of human resources (X1) and physical resources (X2) with progressive instruction (Y) across states (see Table 3): \( Y = 44.7 + .14 (X1 - 47.0) + .09 (X2 - 46.3) \) where 44.7 is the intercept, .14 and .09 are the slopes for X1 and X2 each, 47.0 and 46.3 are the statewide means of X1 and X2. Since the estimated effect of X1 on Y is about 1.5 times greater than the estimated effect of X2 on Y, we may better choose to allocate more resources to enhancing professional development than to upgrading classroom equipment or materials. Suppose that we decide to spend 1.5 times more on X1 than on X2. Substituting \((1.5 \times X2)\) for \(X1\), the above equation is reduced to \( Y = 33.9 + .3 \times X2\). If a desirable level of Y is set around 44.5, the statewide average of progressive instruction measure, then one should expect an average \(X2\) of 35.4 and an average \(X1\) of 53.1 for corresponding adequacy measures of instructional resources. These numbers indicate a very limited level of resource availability according to the item calibration map (see Figure 2). Math teachers in schools providing this average level of progressive instruction would have training in all pedagogical skills but need more training in content knowledge. At the same time, math teachers in such schools may receive some available resources for their instruction but still lack student access to school-owned scientific calculators and computers.

**Discussion**

Clearly, there are limits to the inferences that can be drawn from teachers' self-report on instructional resources and practices. Further, the cross-sectional nature of NAEP data limits causal inferences that can be made about the relationships among school resources, practices and outcomes. The findings may have been confounded by extraneous variables such as socioeconomic and other demographic characteristics of schools that were omitted from the models. The analysis of differences between high and low performing schools in their instructional resources and practices helps us identify the correlates of school performance but does not allow us to determine any causal direction of the relationship. In other words, the central question remains: Do more resources and better practices lead schools to higher performance or do higher performing schools simply draw better teachers and get more resources? Thus, the findings of this research should be interpreted with caution.

Despite the limitations of linking teacher survey to student assessment results, the study explored some strategies to evaluate the effectiveness of instructional resource
allocation and use. Several patterns of instructional resource allocation and use emerged from the HLM analyses of the 1992 NAEP 8th grade math teacher survey and student assessment datasets. First, human and physical resources were weakly related to each other, implying that each measure may tap a somewhat unique aspect of school resources for teaching and learning. Second, the availability of both human and physical resources was positively associated with the level of desirable instructional practices. Generally, the effect of human resources was greater than the effect of physical resources. States that produced more progressive instruction tended to use physical resources more effectively than human resources (i.e., more capital-intensive or labor-saving). Second, the level of desirable instructional practices was positively related to the level of academic achievement across states. But states that performed at a higher level were not necessarily more effective in instructional resource use.

While there were some overall patterns of resource allocation and use across states, it needs to be noted that the relationships among instructional resources, practices and outcomes were found to vary significantly from state to state. This means that setting instructional standards for a certain state to meet desirable levels of academic performance should take into account the effectiveness of its current school resource allocation and use. If states have very tenuous relationships among instructional resources, practices and outcomes and thus are hardly effective in instructional resource allocation and use, simply setting high instructional standards and giving schools more resources (e.g., adequate classroom materials and well-trained teachers) won't necessarily improve outcomes. Much more empirical research is needed in this area to better understand and improve the effectiveness of instructional resource allocation and use.

This study is highly exploratory in determining outcome-driven standards for instructional resources and practices. The empirical method used in this study may lead to overfunding because it relies on data from all schools that produce adequate outcomes, including those that may produce adequate outcomes inefficiently. Nevertheless, it clearly shows that the current school accountability movement across the nation and states with high expectation of student performance requires significantly increased investment in education including teacher training and school funding. The research finding is largely congruent with previous studies using empirical observation of the relationship between school funding and student performance at the district level. Subsequent studies need to look at this relationship more closely at the school and classroom levels beyond the conventional analysis of district-level expenditures per pupil. Longitudinal studies using teacher surveys and student assessments can help identify and track specific areas of teacher needs such as curricular materials and professional development to improve student achievement.

Notes

1. This research was supported in part by the New Scholars Program of American Educational Finance Association and National Center for Education Statistics. Views expressed herein are solely those of the author. An earlier version of this paper was presented at the 1999 annual meeting of American Educational Research Association in Montreal, Canada.

2. The average item difficulty was significantly lower for physical resources measures ($M = 43$) than for human resources ($M = 52$). At the same time, there were significantly greater
variations among items in difficulty for human resources ($SD = 22$) than for physical resources ($SD = 9.5$). This difference in the distribution of item difficulty may be attributable to the large gap between the two measures in the number of items used.

3. If the mean squares misfit statistic is greater than 1.3 or smaller than .75, it signifies that there is significant underfit or overfit. For example 1.3 means that there is 30% more variation in the observed data than the Rasch model predicted. A misfit value (mean square) value of $1 + x$ indicates $100x\%$ more variation between the observed and the model-predicted response patterns than would be expected if the data and the model were perfectly compatible (Bond & Fox, 2001).

4. The performance standard for student-level achievement has a different meaning when we apply it to the aggregate school level. Assuming normal distributions of math achievement within schools, schools' average achievement being at or above the Basic achievement level means that the schools should have at least 50% of their students performing at or above the Basic level. The same logic is applied to the Proficient level.
Appendix

HLM Analysis of NAEP Plausible Values

NAEP used item response theory (IRT) to estimate proficiency scores in math for each individual student. However, these proficiency scores are latent variables conditional on the student's responses to several cognitive and background items and are not directly observed. Because the proficiency scores are estimated, there is some amount of uncertainty or variance associated with them. Thus, rather than having a single observed math score, there is a range or distribution of plausible values for each sampled student's proficiency in NAEP math. Plausible values were developed as a computational approximation to obtain consistent estimates of population characteristics in assessment situations where individuals are administered too few items to allow precise estimates of their ability (see Mislevy, 1991; Mislevy, Johnson, & Muraki, 1992).

In NAEP there are five such plausible values for each sampled student resulting from five random draws from the conditional distribution of proficiency scores for each student. The parameter estimates from the HLM analyses are based on the average parameter estimates from separate HLM analyses of the five plausible values (Raudenbush, Bryk, Cheong, & Congdon, 2000). The HLM parameter estimates that are averaged for this analysis include the Betas (Gammas), the parameter variances (Tau), the reliabilites, the Chi-Square test for the parameter variance being zero, and the probability of that Chi-Square value. The way in which standard error of the averaged Gammas is estimated is described below. The t-value is calculated by dividing the average Gamma by its standard error, and the probability of the t value is estimated from a standard t distribution table.

As evident from the formula below, the use of plausible values in an HLM analysis usually increases the standard errors of the Gamma coefficients, making it harder to identify significant correlates of the outcome variable (Arnold, 1993). The standard error of the Gammas consists of two components – sampling error and measurement error. The following routine provided in the NAEP Data Files User Guide was designed to approximate the component of error variance in the analysis due to the error in measurement and to add it to the sampling error:

\[ \hat{t}_m = \frac{1}{M} \sum_{m=1}^{M} \hat{t}_m \]

Let \( \hat{\Theta}_m \) represent the m th plausible value, where m=1 to M sets of plausible values (in our case M=5). Let \( \hat{t}_m \) represent the parameter estimates based on the m th plausible value. Let \( U_m \) represent the variance of \( \hat{t}_m \), or the sampling error. Five HLM runs are conducted based on each plausible value \( \hat{\Theta}_m \). The parameter estimates from these runs are averaged:
The variance of the parameters from these runs are averaged:

\[ U^* = \frac{1}{M} \sum_{m=1}^{M} U_m \]

The variance of the M estimates \( t_m \) is estimated:

\[ B_m = \frac{\sum_{m=1}^{M} (t_m - \bar{t})^2}{(M-1)} \]

The final estimate of the variance of the parameter estimate is the sum of the two components:

\[ V = U^* + (1+M^{-1}) B_m \]

The square root of this variance is the standard error of the Gamma, and it is used in a standard Student's t formula to evaluate the statistical significance of each Gamma.
References


The Author

JAEKYUNG LEE is an assistant professor of education at University at Buffalo, the State University of New York. He was National Academy of Education/Spencer Postdoctoral Fellow and Principal Investigator of national research projects funded by the U.S. Department of Education and National Science Foundation. His current research focuses on the issues of educational accountability and equity.

Correspondence: <JL224@buffalo.edu>