Using High-Stakes Tests to Derive School-Level Measures of Special Education Efficacy

Ann C. Schulte
Department of Psychology
North Carolina State University

Diane N. Villwock
Chapel Hill-Carrboro City Schools
Chapel Hill, North Carolina

Many states now use large-scale testing programs as a major mechanism in their educational accountability efforts. With the mandated participation of students with disabilities in large-scale testing programs, it is possible to use large-scale testing results to examine special education efficacy at the school level. We used special and general education reading outcomes from 6 elementary schools in 1 school district to compare and contrast measures of a school’s performance over time in serving children with disabilities. Three different methods were contrasted: (a) percentage of students in special education reaching grade-level proficiency in reading each year, (b) percentage of students in special education exceeding expected growth in reading each year, and (c) longitudinal growth and percentage of students reaching grade-level proficiency in 5th grade. We discuss the results in terms of deriving a measure of school special education performance that is both accurate and useful for school personnel.

Until recently, it was not uncommon to exclude children with disabilities from large-scale assessment programs (Elliott, Erickson, Thurlow, & Shriner, 2000; McGrew, Thurlow, & Spiegel, 1993). However, two pieces of national legislation have resulted in marked changes in the extent to which students with disabilities participate in large-scale assessment programs, as well as how these outcomes for these students are considered in assessing school quality. In 1997, the Individuals with Disabilities Education Act (Pub. L. No. 105–17; hereafter, IDEA 97) was amended to require students in special education to participate in state- or district-level assessments or in alternate assessments. Four years later, the reauthorization of the Elementary and Secondary Education Act (also

Requests for reprints should be sent to Ann C. Schulte, Department of Psychology, North Carolina State University, Box 7801, 640 Poe Hall, Raleigh, NC 27695-7801. E-mail: schulte@unity.ncsu.edu
known as the No Child Left Behind Act of 2001; hereafter, NCLB) placed even more emphasis on large-scale assessment for children in special education. This legislation required (a) the disaggregation of large-scale assessment results by specific student groups, including students placed in general education and special education; and (b) individual schools to make adequate yearly progress in improving the achievement of students in these disaggregated groups, including students in special education.

This new emphasis on the outcomes of large-scale assessments for students in special education is the culmination of three national trends. First, there have been longstanding concerns about the efficacy of special education services (Dunn, 1968; Hanushek, Kain, & Rivkin, 1998; Haynes & Jenkins, 1986; Moody, Vaughn, Hughes, & Fischer, 2000; Schulte, Osborne, & Erchul, 1998; Shinn, 1986; Will, 1986; Zigmond et al., 1995). Second, standards-based educational reform accompanied by large-scale assessment of students has become widespread in the United States, with 49 of 50 states having adopted some form of academic standards (Council of Chief State School Officers, 2000). Third, there has been rising concern among policymakers and researchers that the routine exclusion of students with disabilities from large-scale assessments may distort indicators of the efficacy of educational reforms (Allington & McGill-Franzen, 1992; Heubert & Hauser, 1999; McDonnell, McLaughlin, & Morison, 1997; Zlatos, 1994) and limit access of students in special education to this major reform vehicle (Elliott et al., 2000; McDonnell et al., 1997; McLaughlin & Warren, 1992).

Although there have been many concerns about the consequences of excluding children in special education from large-scale educational assessment and accountability programs, very little is known about the consequences of including students with disabilities in high-stakes assessments (Koretz & Barton, 2003; McDonnell et al., 1997; Schulte, Villwock, Whichard, & Stallings, 2001). Students in special education, as a group, present unique characteristics that may interact with particular aspects of large-scale testing and accountability programs, especially when scores for students in special education are disaggregated and reported separately. These interactions can affect the accuracy of conclusions drawn from the test data and the impact of the testing program on students in special education (McDonnell et al., 1997; Ysseldyke & Bielinski, 2002).

For example, Ysseldyke and Bielinski (2002) used state-level data to demonstrate that student special education entrance and exit patterns affect the magnitude of the performance gap observed between general and special education across grades. They tracked a cohort of students as they progressed from fourth to eighth grade using two methods they termed the cohort dynamic and cohort static approaches. With the cohort dynamic approach, mean scores for students in special and general education were reported each year based on their special education status for the particular year in question. With the cohort static approach, mean scores for students in special and general education were reported each year based on students’ special education status in the first year in large-scale testing (fourth grade). When scores were reported using the cohort dynamic approach, the gap between the scores of special and general education students widened considerably as grade level increased. With the cohort static approach, the gap remained relatively constant from fourth to eighth grade. The difference in outcomes between the two reporting methods was the result of higher perform-
ing students in special education reentering general education and lower performing general education students entering special education each year. Ysseldyke and Bielinski suggested that the cohort static approach be used in tracking the progress of students served in special education because it provided a more accurate picture of individuals' response to educational services over time.

In contrast to Ysseldyke and Bielinski (2002) who used state-level data, NCLB requires school-based reporting of disaggregated special education outcomes. However, school-based reporting of special education outcomes poses validity concerns that only partially overlap with concerns raised about reporting results at the state or district level. For example, large-scale testing programs often are designed so that results can be reliably and validly reported at the classroom or grade level (e.g., cumulated across a minimum of 25 students). In many cases, the number of students in special education at a school, particularly if results are disaggregated by special education category or by grade, is much less than this number. The smaller number raises concerns about the reliability of the results reported (Koretz, 1996).

Despite the difficulties of school-based reporting of special education outcomes, there are several reasons why this practice may be desirable. First, such an approach fits well with the trend toward site-based management approaches to accountability in U.S. school districts (Summers & Johnson, 1996). Second, theories of policy implementation suggest that schools are most effective when responsibility and accountability for decision making rests with the staff who are directly responsible for providing instruction (Patterson, Purkey, & Parker, 1986; Weatherly & Lipsky, 1977). Finally, accurate summary indicators of special education efficacy at the school level would allow systematic investigation of the characteristics of schools that are associated with positive outcomes for students with disabilities.

The purpose of this study was to use school results on a large-scale, high-stakes reading test across several years to illustrate the complexities and issues involved in reporting school-based special education outcomes. Test scores from six elementary schools in one school district in North Carolina were used to examine three different ways of reporting school-based reading outcomes for students in special education in terms of their ability to provide meaningful, valid information about a school's functioning with students in special education: (a) percentage of students in special education reaching grade-level proficiency in reading each year, (b) percentage of students in special education exceeding expected growth in reading each year, and (c) longitudinal growth and percentage of students reaching grade-level proficiency in fifth grade. General education results for each school were also used in the analyses to allow comparisons of reporting issues in general versus special education as well as to illustrate problems unique to reporting special education outcomes.

Results from North Carolina's large-scale testing program are particularly useful for comparing different reporting and accountability models because the program allows test results to be considered in several different forms including developmental scores linked across grade levels, grade-level pass rates, and norm-referenced scores. In addition, the state has developed an "expected growth" accountability model that sets growth-based goals for schools at each grade, adjusted for the entering characteristics of students in the grade.
A final advantage of using test data from a North Carolina school district is the overall quality of the testing program. North Carolina's testing and accountability program has been widely studied (e.g., McDonnell & Choisser, 1997), and a recent independent review of the overall quality of state accountability programs ranked North Carolina's school testing and accountability program first in the nation (Princeton Review, 2002). Thus, by studying a district within North Carolina, reporting models for students in special education can be examined within a testing program that represents current best practice for large-scale testing and accountability.

METHOD

State Context

North Carolina's accountability and large-scale testing program has been described in detail elsewhere (Public Schools of North Carolina, 1996a, 1996b; Schulte et al., 2001). At the time this study took place, students in Grades 3 through 8 were tested annually with reading and mathematics tests as well as a writing test in selected grades, which reflected the North Carolina Standard Course of Study (general education curriculum). Across grade levels, forms of each test were linked to a common developmental scale allowing assessment of growth within a specific subject area across multiple years. In judging a school's effectiveness, student annual growth in reading and math was calculated by subtracting a student's score on last year's grade-level test from his or her score on the current grade-level test (in third grade, a fall pretest was used because no group testing took place in second grade). Actual growth was then compared to expected growth, with expected growth calculated using a regression-based formula that adjusted for regression toward the mean and student overall academic proficiency. At the elementary and middle school levels, all teaching staff at a school received cash bonuses if average school growth across reading, mathematics, and writing met or exceeded the growth standards. The amount of bonus pay increased if the school reached a "high-growth" standard exceeding expected growth by 10%.

For students in special education, decisions about participation or nonparticipation in the testing program and the types of accommodations if participating were part of the development of the individualized education program (IEP). However, nonparticipation in the testing program removed a student from the Standard Course of Study and if nonparticipation continued into high school, resulted in ineligibility for a high school diploma. Additional components were added to the accountability model in later years (e.g., the availability of alternative testing so that students with severe disabilities also were included in the accountability program) but are not discussed here because they were not in place at the time of the study.

District Context

At the time of data collection, the school district described herein was among the top performing in the state in terms of student achievement. Located in a university town, the
The district consisted of 11 schools serving approximately 8,000 students at the beginning of the study (7 elementary, 2 middle schools, and 1 high school) and 13 schools serving 8,500 students by the close of the study (7 elementary, 3 middle school, 3 high schools). Data from 6 of the 7 elementary schools in the district were included in data analyses reported here. The 7th school was excluded because it was in its first year of operation at the beginning of the study, and its identification practices, special education services, and staff were in more flux than were the remaining 6 elementary schools. Table 1 presents the demographic characteristics for 6 schools. As is evident from the table, the overall population served by each school was similar.

The district had a strong reputation for its services of students with exceptionalities. A high level of local funding allowed caseloads for special education teachers to be well below the maximum caseloads set by the state as well as below the average caseloads for special education teachers in nearby districts. For example, caseloads for cross-categorical resource teachers at the elementary level in the district ranged from 18 to 25 students. Part-time special education placements were provided at all schools through cross-categorical resource rooms. When students needed special education services more than 60% of the day, they were served in "district-level" classes. Individual schools hosted a portion of these classes that drew students from all elementary schools. This organizational arrangement allowed general education, special education, and related services staff to develop expertise in providing services to students with particular types of needs. School 1 housed a classroom for students with autism. Schools 2 and 4 housed classrooms for students with behavioral/emotional disabilities. School 3 housed a classroom for students with mental retardation, and School 5 housed two classrooms for students with autism and two classrooms for students with mental retardation. School 6 did not host a district-level program.

Participants

Student data regarding general or special education placement and End-of-Grade (EOG) Reading scores were included in this study if a student had attended one of the six elementary schools included in the study at some time during the 1996 to 1997, 1997 to 1998, or 1998 to 1999 school years. Table 2 provides descriptive information for all students who had received special education services at least some time during the time period and grades relevant to the study.

Measures

The primary measure of interest in this study was the North Carolina EOG Test of Reading Comprehension (EOG-Reading). This test consists of approximately 10 reading passages reflecting a variety of reading materials including poetry and literature but with an emphasis on nonfiction. Three to eight questions follow each text passage and require the respondent to read for a variety of purposes (e.g., information, literary interpretation). There are separate versions for each grade level in which the test is administered (three to eight) and alternate equated forms at each grade level. Test results can be reported in a variety of forms, including developmental scale scores (on a single, continuous, equal in-
TABLE 1
Demographic Characteristics by School and Year

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>49</td>
<td>47</td>
<td>47</td>
<td>48</td>
<td>47</td>
<td>50</td>
</tr>
<tr>
<td>Male</td>
<td>51</td>
<td>53</td>
<td>53</td>
<td>52</td>
<td>53</td>
<td>50</td>
</tr>
<tr>
<td>Ethnicity (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>26</td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td>White</td>
<td>75</td>
<td>72</td>
<td>67</td>
<td>61</td>
<td>62</td>
<td>57</td>
</tr>
<tr>
<td>Other</td>
<td>7</td>
<td>9</td>
<td>13</td>
<td>13</td>
<td>16</td>
<td>24</td>
</tr>
<tr>
<td>% free and reduced lunch</td>
<td>18</td>
<td>21</td>
<td>21</td>
<td>32</td>
<td>34</td>
<td>39</td>
</tr>
<tr>
<td>No. of students (all grades)</td>
<td>546</td>
<td>564</td>
<td>570</td>
<td>528</td>
<td>529</td>
<td>546</td>
</tr>
</tbody>
</table>
TABLE 2
Characteristics of Children Served in Special Education in Grades 3 Through 5 From 1997 to 1999

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>208</td>
<td>32.0</td>
</tr>
<tr>
<td>Male</td>
<td>443</td>
<td>68.0</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td>210</td>
<td>32.3</td>
</tr>
<tr>
<td>White</td>
<td>412</td>
<td>63.3</td>
</tr>
<tr>
<td>Other</td>
<td>29</td>
<td>4.4</td>
</tr>
<tr>
<td>Primary exceptionality classification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autism</td>
<td>20</td>
<td>3.1</td>
</tr>
<tr>
<td>Emotional/behavioral disorder</td>
<td>44</td>
<td>6.8</td>
</tr>
<tr>
<td>Learning disability</td>
<td>353</td>
<td>54.2</td>
</tr>
<tr>
<td>Mental retardiation</td>
<td>49</td>
<td>7.5</td>
</tr>
<tr>
<td>Other health impairment</td>
<td>90</td>
<td>13.8</td>
</tr>
<tr>
<td>Speech/language</td>
<td>78</td>
<td>12.0</td>
</tr>
<tr>
<td>Other</td>
<td>17</td>
<td>2.6</td>
</tr>
</tbody>
</table>

Note. N = 651.

terval scale across grade levels), grade-level percentiles, and grade-level proficiency levels. A technical manual is available that provides reliability and validity data (Public Schools of North Carolina, 1996a) as well as normative data. Internal consistency reliability estimates are above .90 at all grade levels, and the standard error of measurement ranges from 2 to 5 points for an individual’s test score depending on the grade and proficiency level of the student. The higher standard errors are associated with the more extreme scores at each grade level. The fifth-grade-level EOG-Reading has been correlated with the reading portion of the Iowa Test of Basic Skills. With three separate samples of 3,000 students, the concurrent validity estimates ranged from .80 to .84. On the 1992 to 1993 state-level administration of the test, developmental scale score means ranged from 142.9 (third grade) to 158.7 (eighth grade), with standard deviations from 8.9 to 9.9.

In addition to the EOG-Reading in Grades 3 through 8, a reading pretest for the third grade was administered during the first 3 weeks of third grade. This test is shorter than the EOG-Reading and assesses children’s mastery of the second-grade curriculum in reading. It was developed to allow the assessment of growth from the beginning to end of third grade.

Grade-level proficiency standards for the EOGs were established by asking over 5,000 teachers to categorize their students into one of four achievement levels ("below basic," "basic," "proficient," and "advanced") and then using this information as the basis for setting cut scores for four proficiency levels (Public Schools of North Carolina, 1996a). Students are considered to have reached grade-level proficiency in an academic area if their scores fall in the proficient or advanced range for their grade.

In addition to the EOG-Reading, an EOG-Mathematics and third-grade mathematics pretest were part of the state’s accountability program. The tests were similar in form and test characteristics to the EOG-Reading and are described in more detail in Public
Schools of North Carolina (1996a) and Schulme et al. (2001). They are mentioned here because they were used in estimating overall academic proficiency in the calculation of expected growth.

Procedure

**EOG test administration.** The EOG tests were administered by the school district in May of each year of data collection. For general education students, classroom teachers administered the tests in the general education classroom. As a test security measure, three different versions of the test were available each year for each grade level, with the particular version received randomized by child within classroom. The EOG-Reading was administered on one day, with the EOG-Mathematics administered on a second day in the same week. Both tests were administered in 2-hr morning sessions.

Children in special education at the time of each yearly testing session were eligible to receive a range of accommodations that had been approved by the state. The specific accommodations each child received were determined by the student’s IEP committee, with the expectation that the same accommodations used in the testing were provided for daily work in the classroom.

**Determination of special education status.** Special education status was determined by consulting the April special education head count data provided to the North Carolina Department of Public Instruction by the district in each year of the study. As these were the special education data certified as correct by the superintendent and subject to audit by the state, they were considered the most accurate representation of demographic information and special education status available. Students appearing in the April head count for whom no reading test scores were included in the district’s records were counted as nonparticipants in the testing program.

**RESULTS AND DISCUSSION**

Preliminary Data Analyses

**Pretest-posttest correlations.** For the entire sample of students in special education participating in large-scale testing, we calculated the correlation between yearly pretest and posttest scores at each grade level. The following correlations were obtained:

- Third-grade reading pretest and EOG-Reading in third grade, $r(267) = .72, p < .0001$;
- EOG-Reading third grade and EOG-Reading fourth grade, $r(249) = .81, p < .0001$; and
- EOG-Reading fourth grade and EOG-Reading fifth grade, $r(243) = .81, p < .0001$.

These figures indicate a substantial portion of the variance in schools’ annual outcomes for students in special education is determined by the entering achievement level of the students. The smaller correlation between the third-grade pretest and EOG-Reading likely reflects attenuation due to the shorter length and lower reliability of the third-grade pretest.

**School-based special education prevalence rates and testing participation rates.** Table 3 describes the special education population in Grades 3 through 5 in each of the six schools included in the study. Several aspects of the table are notable. First, the
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Autism</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>E/BD</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>6</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>LD</td>
<td>27</td>
<td>36</td>
<td>34</td>
<td>27</td>
<td>30</td>
<td>27</td>
<td>12</td>
<td>10</td>
<td>16</td>
<td>23</td>
<td>16</td>
<td>16</td>
<td>26</td>
<td>39</td>
<td>33</td>
</tr>
<tr>
<td>MR</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>10</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>OHI</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>6</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>Sp/lang</td>
<td>6</td>
<td>7</td>
<td>10</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>All SE</td>
<td>45</td>
<td>51</td>
<td>60</td>
<td>43</td>
<td>44</td>
<td>49</td>
<td>24</td>
<td>20</td>
<td>31</td>
<td>43</td>
<td>41</td>
<td>48</td>
<td>51</td>
<td>66</td>
<td>64</td>
</tr>
<tr>
<td>% SE</td>
<td>15.5</td>
<td>16.9</td>
<td>19.0</td>
<td>16.9</td>
<td>14.0</td>
<td>17.3</td>
<td>11.9</td>
<td>10.6</td>
<td>12.7</td>
<td>14.1</td>
<td>14.0</td>
<td>16.8</td>
<td>16.2</td>
<td>19.1</td>
<td>19.8</td>
</tr>
<tr>
<td>% tested</td>
<td>84.4</td>
<td>100</td>
<td>88.3</td>
<td>90.7</td>
<td>100</td>
<td>95.9</td>
<td>66.7</td>
<td>75.0</td>
<td>77.4</td>
<td>95.3</td>
<td>92.7</td>
<td>93.8</td>
<td>68.7</td>
<td>77.3</td>
<td>81.2</td>
</tr>
<tr>
<td>% SE, excluding district level</td>
<td>13.1</td>
<td>16.9</td>
<td>17.6</td>
<td>15.6</td>
<td>12.1</td>
<td>16.4</td>
<td>8.7</td>
<td>8.2</td>
<td>11.6</td>
<td>12.8</td>
<td>13.7</td>
<td>15.9</td>
<td>12.9</td>
<td>17.0</td>
<td>16.7</td>
</tr>
<tr>
<td>% tested, excluding district level</td>
<td>94.9</td>
<td>100</td>
<td>96.4</td>
<td>92.3</td>
<td>100</td>
<td>100</td>
<td>94.1</td>
<td>100</td>
<td>85.7</td>
<td>100</td>
<td>95</td>
<td>100</td>
<td>87.2</td>
<td>89.5</td>
<td>100</td>
</tr>
</tbody>
</table>

Note. E/BD = emotional/behavioral disorder; LD = learning disability; MR = mental retardation; OHI = other health impairment; Sp/lang = speech/language; SE = special education.
population served in special education at each school shifted somewhat from year to year. For example, in 1997, 6.7% (3 students) of School 1’s special education population were students with autism, and 60% (27 students) were students with learning disabilities. In the following year, all students with autism in Grades 3 through 5 had aged out of the school or moved, and students with learning disabilities represented 71% of the students in special education. Such fluctuations in a relatively small sample might be expected to affect both the percentage of students participating in large-scale testing (as is evident in the fluctuations in percentage of students tested for School 1 from 1997 to 1998) and/or the achievement outcomes reported for the school for those years. Second, the special education prevalence rate varied by school. Some of the differences among schools were likely due to the presence of specialized district-level programs at particular schools. However, even when students in district-level classes are excluded from estimates of special education prevalence, there appeared to be differences in special education identification patterns across schools. Such a finding is consistent with previous research showing low reliability in special education classification decisions for high-incidence disabilities (Shepard, 1989) and local variation in special education classification practices (Singer, Palfrey, Butler, & Walker, 1989). Third, there was a general trend of increasing special education identification, with five of the six schools serving a higher percentage of students in special education in 1999 than 1997. Such variations in special education classification practices by school or region or across time within one school make comparisons of outcomes across time or different schools problematic (Koretz & Barton, 2003; Schulte et al., 2001).

**Percentage of overlap in students tested from year to year.** One difficulty in reporting annual school testing outcomes aggregated across grade levels is that each year approximately one third of students exit the school (fifth graders) and one third age into the testing program (i.e., enter the grade where the testing program begins). When combined with family moves in and out of a school’s catchment area and special education entrances and exits, the population of students served in special education each year at a school may vary considerably, as illustrated in Table 4. Unless the cohorts exiting and entering are quite similar, differences in outcomes across years may be largely due to entry characteristics of the cohort served rather than a reflection of the quality of the services provided by the school staff (Koretz, 1996; Willms & Raudenbush, 1989). Differences in cohorts are more likely to affect outcomes when sample sizes are small, student ability levels are diverse, and student achievement level affects the probability of being included or exited from the cohort (Koretz, 1996; Meyer, 1995; Ysseldyke & Bielinski, 2002).

**Percentage of Students Reaching Grade-Level Proficiency**

Figure 1 presents the percentage of students in Grades 3 through 5 reaching grade-level proficiency in general education and special education. Special or general education status was determined on an annual basis for each child. Perhaps the most striking aspect of Figure 1 is the variability across schools in outcomes for students in special education in comparison to the variability in outcomes for students in general education. For all but
one of the schools, 90% of students in general education scored at or above grade level each year. In contrast, the percentage of students in special education reaching grade-level proficiency at each school ranged from below 40% to over 80% in the 1st year, from 45% to 86% in the 2nd year, and from 54% to 79% in the 3rd year.

There are likely three reasons for the low variability observed in the percentage of general education students reaching proficiency across schools. First, fluctuations in achievement could have been masked by ceiling effects for general education students because only changes in test scores that would result in changes in the total number of students reaching grade-level proficiency would be apparent with this reporting metric. Second, the general education population at each school was relatively large and stable in terms of student characteristics (see Table 2). Therefore, cohort differences in characteristics related to student outcomes (e.g., achievement level at the beginning of the year) were unlikely to result in marked fluctuations in the percentage of students reaching proficiency. Third, a portion of the students in general education encountering difficulty each year were referred for, and then placed in, special education.

There was considerable variability in special education outcomes between schools and across years. In particular, School 3, which had the smallest number of students in special education across the 3 years (see Table 2), showed a drop in percentage of students proficient from the 1st to the 2nd year—from 84% to 47%. Across the 3 years, there was no overlap in the performance of School 5, the highest performing school, and four of the five remaining schools. School 2 was ranked the lowest in 1997 and 1998, and fifth out of six in 1999.

**Percentage Achieving High Growth**

As noted earlier, a unique feature of North Carolina's testing and accountability program is the use of expected growth standards in determining school performance, with growth assessed by comparing simple gain scores (student EOG score for the present grade mi-
FIGURE 1 Percentage of students in special and general education achieving grade-level proficiency in reading by school.

With his or her EOG score for the previous grade) to an expected progress standard that is adjusted for regression toward the mean and overall academic proficiency (Public Schools of North Carolina, 1996b).

As an alternative to percent proficient as a measure of school efficacy with students in special education, we calculated the percentage of students in special education each year achieving high growth (110% of expected growth) in reading by school. We selected
percent achieving high growth rather than expected growth as a metric for examining school special education efficacy because as a group, students with special education tended to score below grade-level expectations. Only by exceeding growth expectations could students in special education close the gap between their performance and grade-level expectations.

The percentage of students in special education at each school achieving high growth each year along with the percentage of students in general education achieving this same standard are presented in Figure 2. In contrast to Figure 1, school outcomes for students in special and general education were less discrepant when growth was considered rather than grade-level proficiency. Special education outcomes for the six schools were quite similar when growth was considered. In contrast to the results using percent proficient in which School 2 performed less well than the other schools, School 2 was the only school to produce high growth in over 50% of its students in special education all 3 years.

Longitudinal Results by School

A final option in reporting special education results by school is an examination of longitudinal results by student across their elementary school years. Variance in school annual outcomes across multiple years inevitably reflects a mix of variance due to differences in cohorts of students and variance due to changes in individual student's academic achievement (Koretz, 1996; Richards, 1975). In contrast, cumulating results longitudinally by student provides information about a question of central importance in serving students with disabilities (Ysseldyke & Bielinski, 2002)—Did individual students with disabilities make substantial progress in academic areas over time?

In calculating longitudinal results by student, we examined school performance with students served in special education from third grade (when the North Carolina testing program began) to their exit from elementary school in fifth grade. A potential disadvantage of this metric is that there are only a small number of students in each school who have been served in special education who exit the school each year. To address this issue, we cumulated results across three cohorts of exiting fifth graders (i.e., children who were fifth graders in 1997, 1998, and 1999). We also included any child who had ever been served in special education in Grades 3 through 5 regardless of their special education status in fifth grade. This is a variant of the Ysseldyke and Bielinski’s (2002) cohort static reporting method that assures that students who improved sufficiently to be exited into general education are included in assessing special education outcomes. However, unlike the cohort static method, it also includes the substantial number of students identified for special education beyond the first year of high-stakes testing in estimating special education outcomes. Although this method may misrepresent the performance of special education teachers (because some students included received no special education services in some years), it can be considered a measure of whole school functioning relative to students with disabilities because it reflects the school’s performance with students in special education regardless of where they were served.

Figure 3 presents the longitudinal outcomes by school cumulated across three cohorts of students who had exited the school. In contrast to the reporting methods that were cross sectional (Figures 1–3), results presented longitudinally by student show steady growth
across grades at all schools. As is evident from the figure, there were substantial differences among schools in fifth-grade outcomes, $F(5, 220) = 4.41, p < .001$. However, when third-grade EOG-Reading was used as a covariate, the differences between schools were no longer significant, $F(5, 196) = 0.93, p < .47$. Third-grade reading was a significant covariate, $F(1, 196) = 188.89, p < .0001$, accounting for 44.7% of the corrected total variance, whereas school accounted for 1.1% of the corrected total variance.¹

¹Only students who had attended the same school for Grades 4 and 5 were included in these analyses. Sample sizes vary between the two analyses due to missing third-grade reading achievement scores for some students.
FIGURE 3 Longitudinal reading growth means and percentage of students attaining reading proficiency in Grade 5 by school for children who received special education services.

This finding suggests that either schools’ effects were operative prior to third grade, or among this sample of six schools, there was very little variability in the quality of services for students in special education progressing through the general education curriculum.

GENERAL DISCUSSION

“Creating a fair and responsible reporting mechanism is one of the major challenges associated with the participation of the students with disabilities in large-scale assessments and public accountability systems” (McDonnell et al., 1997, p. 184). With individual schools now a primary focus of accountability and school improvement efforts because of NCLB (2001), this study examined options for presenting information about the performance of individual schools relative to the students in special education they serve. Disaggregated special education outcomes in reading from a large-scale testing program were presented for six schools using three reporting metrics, including percentage of students achieving grade-level proficiency, percentage making expected growth, and longitudinal outcomes and fifth grade reading proficiency for all students served in special ed-
ucation across three cohorts. The results illustrate a number of important considerations in developing an accurate, fair, and useful measure of a school’s effectiveness with children served in special education. In this section, we identify critical issues in the development of a meaningful school-based reporting mechanism that reflects schools’ efficacy in serving students with disabilities.

First, schools appeared to differ in the special education populations they serve, with two major sources of variation being the location of district-level classrooms for students with more severe disabilities and individual school’s special education identification practices. In this study, the low participation rate of children in self-contained classrooms in the large-scale testing meant that the presence of district-level classrooms affected a school’s participation rate rather than its achievement outcomes. However, with more stringent participation criteria and the availability of alternative testing formats, the presence of district-level classes could also exert an impact on achievement outcomes, particularly for schools that host specialized programs for students with severe cognitive impairments.

Without consideration of the impact of specialized programs on schools’ achievement outcomes for students, there may be disincentives for schools to host such programs. Reassigning students, or the scores for students, with severe cognitive disabilities back to their “home schools” are two options to deal with this issue. However, neither strategy will successfully address the issue that some schools and districts may serve high proportions of children with severe cognitive disabilities because families have relocated to their catchment areas to access services provided by the school or community.

Another source of variation in special education populations by school appeared to be differences in identification practices. The percentage of children served in special education in each school varied (see Table 1) as did the mean achievement of students when large-scale testing began in the third grade (see Figure 1). Differences in the stringency of school identification criteria may affect the overall percentage of students in special education achieving at grade level. Schools with more stringent identification criteria for special education may have a more severely disabled population and lower achievement as a result.

A second consideration in developing a fair reporting mechanism for schools’ special education performance is the variation in school special education population from year to year. Schools’ special education population varied as a function of grade-level turnover and entry and exit patterns for special education. The issue of how to separate variance in school achievement due to cohort differences from variance due to the effectiveness of school staff is an issue that has been extensively studied in educational measurement (Hanushek & Taylor, 1990; Meyer, 1995, 1996; Raudenbush & Bryk, 1986). Relative to this issue, Meyer (1995) has distinguished between two uses of large-scale test results. Test results may be used descriptively to provide information about the achievement of particular groups of students or as performance indicators to provide a measure of how well the school serves a particular group of students.

Many measures of school achievement that are useful as descriptive measures, such as school means or the percentage of students reaching grade-level proficiency, are inaccurate performance indicators (Hanushek & Taylor, 1990; Koretz, 1996; Meyer, 1995, 1996; Willms & Raudenbush, 1989). This inaccuracy arises because factors unrelated to
the school's role in student performance in a particular year, such as the entry characteristics of the cohort, produce variations in test scores from year to year that can mask the effect of the quality of services provided by school staff on achievement (Koretz, 1996; Willms & Raudenbush, 1989).

As a result of this problem, several researchers (Hanushek & Taylor, 1990; Meyer, 1995, 1996; Raudenbush & Bryk, 1986) have argued that "value-added" measures should be used rather than more common descriptive measures in assessing school performance. Value-added measures use statistical regression to estimate and correct school performance measures for factors that contribute to student outcomes but are unrelated to the quality of education provided by the school. As such, value-added measures are more valid indicators of school performance and quality. When descriptive measures and value-added measures of school performance yield conflicting results, the use of descriptive measures could result in actions that harm students, such as discontinuing an effective program or continuing an ineffective program (Meyer, 1995).

Value-added reporting metrics vary markedly in the number of factors extraneous to a school's contribution to student achievement for which they attempt to adjust. In a study of the bias and error resulting from using descriptive outcomes versus different value-added models in judging states' contributions to student achievement, Hanushek and Taylor (1990) estimated that error variance was 7 to 12 times as large as the underlying variation in school quality. Simply reporting descriptive test results overestimated the effect of schools on student achievement by a factor of 3 to 5. A simple value-added model adjusting performance indicators only for student achievement level at entry markedly reduced bias and error in estimating a schools' affect on student achievement, although more complex models reduced bias even more. Other researchers' have confirmed that a substantial amount of the error in assessing school staff performance can be eliminated by simply adjusting school performance estimates for student achievement at entry (Meyer, 1996; Richards, 1975).

In this study, two of the reporting mechanisms could be considered value-added measures. The percentage of students achieving high growth (Figure 2) adjusts for student achievement level in the previous grade as well as student overall academic proficiency. In addition, the longitudinal analysis (Figure 3) included a test of whether significant differences in achievement in fifth grade across schools could be explained by factors beyond students' third-grade achievement status.

Choosing a Reporting and Accountability Mechanism

Measures. What measure or set of measures constitutes a "fair and responsible" reporting mechanism for special education? This study, although limited by its use of only six schools in a single district, illustrated some of the complexities in reporting and interpreting special education outcomes for individual schools. Whether judging schools more or less effective with children who have special needs or determining whether an individual school is improving in its efforts to enhance the achievement for all children, the reporting and interpreting of the results of large-scale tests for children served in special education programs is complicated by multiple factors. Some of these are common to the interpretation of school effects for all children (e.g., cohort differences, confound-
ing of preexisting differences with current school effects in many measures); others are unique to special education (e.g., bias in outcomes because of entry and exit patterns).

In our judgment, a suitable reporting mechanism for school-based special education outcomes should (a) primarily reflect the contribution of the school staff to children's achievement for the time period assessed (i.e., be a valid performance indicator), (b) provide information about students' performance relative to expectations set for them, and (c) provide school staff with both annual information about students' outcomes because of its formative evaluation value and longitudinal information because of its greater reliability. Based on these criteria, we suggest that annual information about the percentage of students in special education achieving high growth and longitudinal information about students' reading growth across grades and proficiency at fifth grade (cumulated across cohorts) represent a strong combination of measures for examining school-based outcomes for children with disabilities. This combination provides annual value-added information for children served in special education that year and information about progress through the grades and fifth-grade exit status for all children ever served in special education. Combined with special education prevalence information, these two measures provide substantial information about the quality of the schools' special education services in the context of the population served.

**Accountability standards.** Given these two measures as the reporting mechanism, what might constitute an appropriate set of standards on which to judge a school's special education services effective?

Within the context of the schools examined in this study, a criterion of over 50% of students in special education achieving high growth annually seems to be a reasonable yet high standard for judging special education services effective. Such a standard would assure that achievement growth for the majority of students in special education is sufficient to accelerate growth toward grade-level standards. This standard, combined with the expectation that a high percentage of students served in special education (e.g., 75%) exit the school functioning on grade level, represents a reporting and accountability mechanism that provides incentives for schools to provide high-quality special education, orients the staff toward students meeting grade-level standards, and takes into account the serious measurement issues that are posed in assessing school outcomes for students in special education.

**Future Directions**

In this study, we examined a small sample of schools and a limited number of reporting mechanisms for special education outcomes. However, it is a first step in examining the characteristics of different options for reporting school-level outcomes on large-scale tests for students in special education, and specifying the criteria for identifying "good" schools for children with disabilities. Most previous work discussing suitable performance standards for special education has not used results from large-scale assessments as a primary outcome (e.g., Deno, Fuchs, Marston, & Shin, 2001; Stone & Doane, 2001).

In future studies, larger numbers of schools will assure that results are generalizable to a broad range of schools. A large sample of schools would also allow an examination of the...
relation between various service delivery features and special education outcomes (e.g., inclusionary programs, direct and strategy-based instruction) and an examination of the correlations among various reporting metrics. Knowledge of the correlations among reporting metrics would allow an examination of the extent to which simpler measures of school efficacy approximate the validity of more complex reporting mechanisms.

Schools are under increased pressure to improve their outcomes for students with disabilities. However, efforts to improve outcomes for children will be encumbered if the information that we give schools about the quality of their services provides information that is not meaningful, or at worst, is misleading. A careful consideration of school-based assessment outcomes on large-scale tests for students in special education, when tempered with a clear understanding of the sources of bias in these results, offers promise for providing a greater understanding of factors important in achieving high-quality special education services and providing meaningful information to school staff about the services they provide to children with disabilities.

REFERENCES


Copyright of Exceptionality is the property of Lawrence Erlbaum Associates and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.