DOCUMENT RESUME

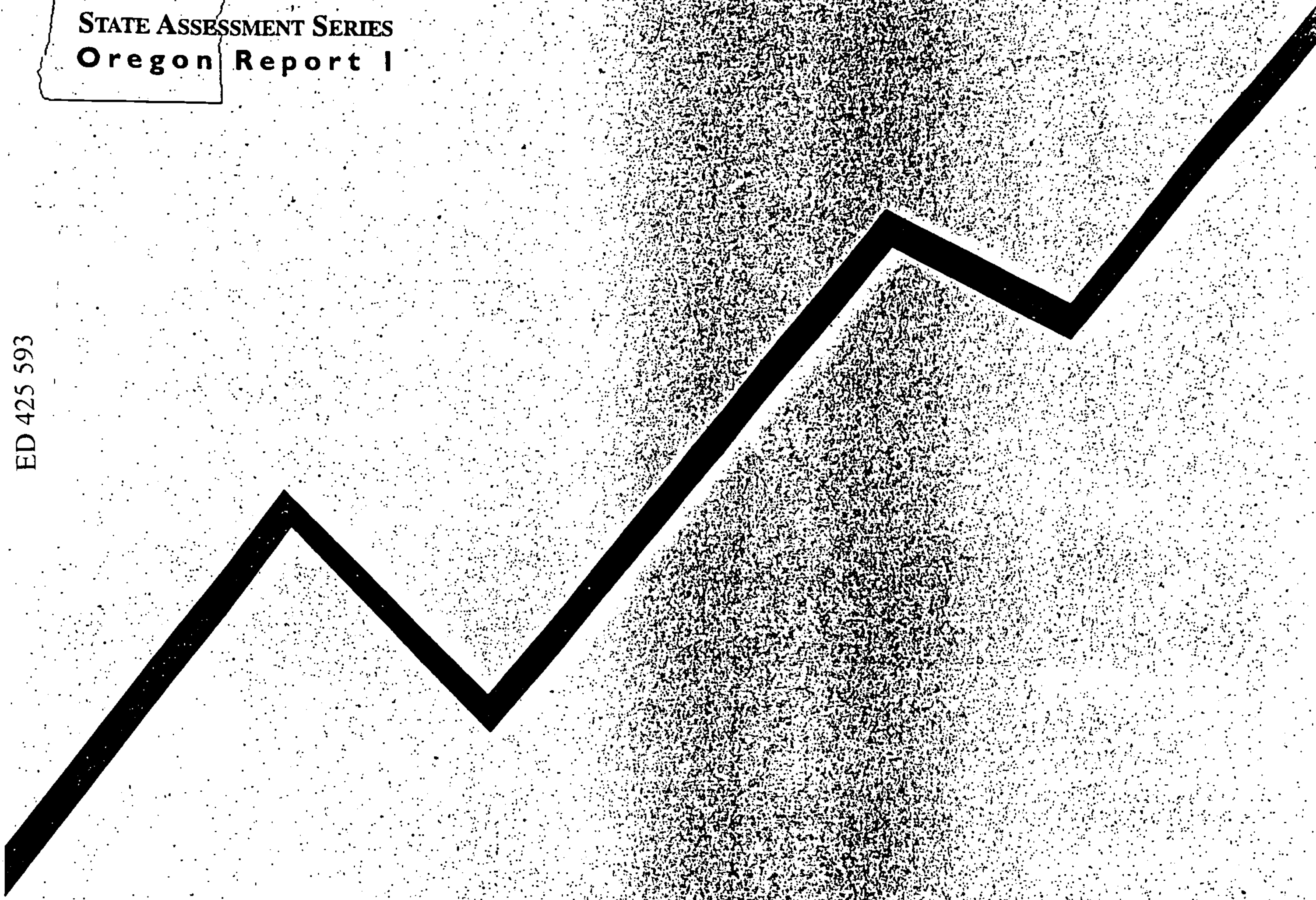
ED 425 593 EC 306 934

AUTHOR	Almond, Patricia; Tindal, Gerald; Stieber, Steve
TITLE	Linking Inclusion to Conclusions: An Empirical Study of Participation of Students with Disabilities in Statewide Testing Programs. State Assessment Series, Oregon Report 1.
INSTITUTION	National Center on Educational Outcomes, Minneapolis, MN.; Oregon State Dept. of Education, Salem.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	1997-10-00
NOTE	40p.
CONTRACT	R279A50027
AVAILABLE FROM	National Center on Education Outcomes, University of Minnesota, 350 Elliott Hall, 75 East River Road, Minneapolis, MN 55455; Tel: 612-624-8561; Fax: 612-624-0879; Web site: http://www.coled.umn.edu/NCEO ($10).
PUB TYPE	Reports - Research (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Accountability; *Data Collection; *Disabilities; *Educational Assessment; Elementary Secondary Education; Input Output Analysis; *Outcomes of Education; *Recordkeeping; Research Problems; School Effectiveness; State Programs; *Student Participation
IDENTIFIERS	*Oregon

ABSTRACT

This report discusses the findings of a study that attempted to investigate the participation rates of students with disabilities in the Oregon statewide testing program. The study joined individual student data from the Oregon Statewide Assessment Program and the Oregon Special Education Child Count to examine participation rates. The report discusses the problems encountered in merging two extant data sets and highlights these findings as relevant for any state department with a testing program that creates and uses a database separate from the special education child count files. The study found an unsettling level of disagreement between various indicators of special education membership. The data collections are operated by different offices under separate statues and federal regulations, and data are maintained in separate unrelated files. Recommendations for examining the capability of the assessment program to assess and report on all of its students adequately are provided, including: (1) examine various data collections involved; (2) conduct an analysis of how statues affect which data; (3) include demographic information on testing answer sheets; (4) investigate the circumstances surrounding the marking of exclusion codes; and (5) improve the ability to both aggregate and disaggregate summaries. (Contains 16 references.) (CR)

STATE ASSESSMENT SERIES
Oregon Report 1



Linking Inclusion to Conclusions: An Empirical Study of Participation of Students with Disabilities in Statewide Testing Programs

STATE ASSESSMENT SERIES
Oregon Report 1

Linking Inclusion to Conclusions: An Empirical Study of Participation of Students with Disabilities in Statewide Testing Programs

Oregon Assessment Development and Evaluation Project

Prepared By:

Patricia Almond
Oregon Department of Education

Gerald Tindal • Steve Stieber
University of Oregon

October, 1997

Oregon's School Improvement Plan calls for systemic education reform to raise content and student performance standards throughout the state. This project is developing and field testing science assessments and evaluating current statewide assessments in reading, writing, and mathematics. These high quality statewide assessments are aligned with the challenging State content standards. Assessments are being adapted and modified for limited English proficient (LEP) students and students with disabilities to ensure that all students participate.

This project is supported, in part, by a grant to the Oregon State Department of Education from the U.S. Department of Education, Office of Educational Research and Improvement (Grant #R279A50027). Opinions expressed herein do not necessarily reflect those of the U.S. Department of Education or Offices within it.

NATIONAL CENTER ON EDUCATIONAL OUTCOMES
University of Minnesota • 350 Elliott Hall
75 East River Road • Minneapolis, MN 55455
612.624.8561 • Fax 612.624.0879
http://www.coled.umn.edu/NCEO

This document is available in alternative formats upon request.

Overview

In this study, we investigate the participation rates of students with disabilities in a statewide testing program. We begin the study by following the work of the National Center on Educational Outcomes (NCEO), identifying many of the reasons proffered for excluding students in these assessments. Although participation rates may be a function of poor data collection at the time of testing, we found such problems only begin to tell the story. Indeed, the entire data collection system, from encoding student demographic information at the time of testing to the merging of files using key marker variables, is fraught with problems. In the end, many students with disabilities are lost along the way. We highlight the findings as relevant for any state department with a testing program that creates and uses a database separate from the special education child count files.

Through the work of NCEO, a number of issues have been identified in large scale testing programs. In this brief chronology of the empirical and conceptual literature currently published through NCEO, we highlight two important issues. First, including students with disabilities in large-scale, statewide testing programs is based on many subtle assumptions and distinctions, which, if not made explicit, are likely to render few changes in either policy or practice. Second, inclusion is not mere physical presence, but must be considered in the entire process from scheduling and implementing tests to preparing data files and reporting outcomes; in all of these steps, students with disabilities are "lost" along the way. The purpose of this study is to highlight the issues raised by NCEO and extend them into a practical and operational consideration in the field.

In the early work of NCEO, Ysseldyke and Thurlow (1993) presented a range of views on inclusion in assessment, citing perspectives of several significant writers in the special education literature. Importantly, they noted 13 critical areas to be addressed in both including and accommodating students with disabilities in assessment programs: (a) definitions, (b) data quality, (c) equity, (d) sampling methodology, (e) data aggregation, (f) test standardization, (g) cross-sectional versus longitudinal assessment, (h) instrument adaptation, (i) validity, (j) reliability, (k) range of items, (l) out-of-level testing, and (m) feasibility of special studies. In later work, many of these issues were addressed further, while others were postponed. In this article, we present data on three of these areas, each of which is critical to the integrity of the entire testing-reporting process:

1. Standardization of testing is critical not only for ensuring comparability of test outcomes, allowing student scores to be aggregated, but for ensuring that complete records are part of the data file. Based on our data, we believe most states are likely to have problems with this issue.

2. Data quality is essential for making valid inferences; not only are reports of participation rates a function of this quality, but the inferences that can be made from the achievement scores are likewise related to the quality of the data.

3. Data aggregation is not even worth considering unless the data are collected in a standardized manner, the data files are of high quality, and certain marker variables are part of the data collection process.

Thurlow and Ysseldyke (1993) also raised a question about the national educational agenda: "Can 'all' ever really mean 'all' in defining and assessing student outcomes?" They pointed out that the rhetoric about all students sounds good and inclusive, but that few proponents of all students were dealing with the implications of the inclusive language. Challenges included the following:

- Measurement groups repeatedly developed lists of reasons why it was very difficult to accommodate students with disabilities in state and national testing programs.

- Given the current framework within which most educators (particularly educational administrators) typically operate, it seems easier not to include students with disabilities when thinking about educational outcomes. In the past, many had a separate curriculum.

- The measurement outcomes within existing general education assessment tools frequently were difficult when students with disabilities were included. Generally an adaptation was needed.

In concluding, they pointed to 25-30% of students (not only students with disabilities) for whom higher standards, world class standards, and other reforms raise large questions.

We describe highlights of the current literature on inclusion in assessment as a context for placing our study; we then focus on the three assertions raised above, describing a methodology for studying them that, we believe, is endemic to any state that operates both special education child count files and statewide assessment files. We begin with Algozzine's (1993) perspective:

> To improve assessment outcomes in America's schools, professionals should avoid any practices that produce, encourage, foster, or facilitate separation among student groups. All students should be expected to take all tests and any modifications permitted for any assessment procedures should be permitted for all tests, all assessment procedures, and *all students*. (p. 9)

Statewide Practices on Inclusion in Assessments

A report by Ysseldyke, Thurlow, McGrew, and Shriner (1994) described four purposes for which statewide assessments can be used: (a) to make decisions about student competence, (b) to provide data to inform policy; (c) to compare local educational agencies, and (d) to provide accountability data on criterion-referenced achievement levels. Yet, in an earlier report by Ysseldyke, Thurlow, McGrew, and Vanderwood (1994), it had been noted that "there are differential participation rates across states" (p. 2) and that the factors leading to exclusion consisted of vague guidelines, inconsistently implemented and monitored guidelines, differential sampling plans (of students), unwillingness to make accommodations, altruistic motivation to lessen student distress, and presence of (dis)incentives in reporting outcomes. While they found exclusion rates of 0% to 100%, they also noted that "we do not have a good understanding of the magnitude of exclusion of students with disabilities in state assessment programs" (p. 4). They recommended that students with disabilities need to be included in reporting results, suggesting that the data be aggregated and disaggregated, depending on whether a student received an accommodation within an assessment or received an alternative assessment. In the end, they estimated that as many as 85% of the nearly 5 million students receiving special education services can take a statewide test, often with minor or no accommodations.

Thurlow, Scott, and Ysseldyke (1995) more recently reported that 24 states describe what they do with data on students with disabilities, including reporting results of standard, accommodated, or alternate administrations and reporting the "records of the numbers and who was excluded from the assessment" (p. 6). Finally, in one of their most recent papers, Erickson, Thurlow, and Ysseldyke (1996) pushed the issue of inclusion toward an operational perspective, addressing it from a reporting integrity perspective. They identified several problematic issues that lurk in the background of any attempts to report participation rates, three of which include (a) neglected numerators, which occur from not knowing which special education students actually participate in the assessment; (b) drifting denominators, which occur from differences in the way educators define who is eligible (e.g., all students in any of the 13 federal categories, all students in the public schools, or all students with at least 50% of their time spent in general education classrooms); and finally, (c) fractured fractions, a problem occurring because many statewide test files are collected and collated differently (in time and by personnel) than the special education child count files. They recommended that educators be explicit in the policies and be clear on the practices. In the end, Elliott, Thurlow, and Ysseldyke (1996) presented several checklists that statewide assessment and local education agency personnel can use to maximize the participation of students with disabilities in large-scale assessments.

Most recently, Thurlow (1997) reviewed the statewide assessment practices of two states, Kentucky and Maryland, which stand out among other states in their inclusive practices in calculating participation rates, setting policies about accommodations, aggregating scores, and reporting of the results of assessments. Both state systems reflect a premise that all students count and that accountability must encompass all students. Both systems generally view accommodations as appropriate for support of students with disabilities, and they both essentially assign zero scores when students are kept out of the assessment. Thurlow's review described the policies and reporting systems for the two states based on an analysis of actual state reports and discussions with local school district and state education personnel. Kentucky expects that no more than 2% of the student population will be designated for the alternate portfolio system. Exceeding this percentage triggers an audit. Policies and reporting practices were addressed by Thurlow, but audit data on each state's actual test participation for students with disabilities were not provided.

These efforts by NCEO are an excellent starting point for replacing an unsystematic and implicit system for statewide testing data collection with one that is systematic and explicit. In this study, we extend the operational issues of clarifying the fractured fractions problem by clearly describing the steps that need to be taken to merge two data files. Implicit in this analysis is the need to continue clarifying both the numerator (who actually takes the test) and the denominator (relative to what special education population). We focus on the initial challenge of locating all students when using extant data, the problems inherent in actually merging two separate data files, the challenges encountered in identifying special education students having test scores, and the problems encountered in reporting on the performance of special education students as a subcategory of the total testing population.

Oregon maintains annual electronic data files on both students taking the statewide assessments and students counted on the annual December 1, special education child count. Oregon's legislation for the 21st century insists on high standards and accountability for *all* students. Oregon generally considers accommodations in test administration to be standard and includes scores obtained with accommodations when reporting and aggregating data. In Oregon, some students are exempt from taking individual tests. Exemptions fall into two categories: limited English proficiency and special education. Other students take the test with significant modifications that change the content of the test itself. Scores obtained under modified conditions are not included in aggregating and reporting testing results.

Methodological Issues in Extending NCEO Analyses

We encountered two methodological issues in extending the NCEO analyses. First we wished to establish a preliminary estimate of what to expect. How many students were enrolled? Did all enrolled students take the test? How many were special education students? How would we determine whether we had accounted for all students? In addition, we faced technical challenges in joining student records from two unique data sets (extant data) collected at different points in time, by different offices, and for different purposes.

Establishing an Estimate

The initial issue that we faced involved establishing preliminary estimates from published reports produced from available data sources. We were employing existing data collections for our analysis, but given questions raised about who is included in "all," we wanted to begin with preliminary estimates. We wanted to be able to understand our findings in a context. We compared three counts: Oregon population estimates (Wineberg, 1997), the report of average daily membership-resident (Oregon Department of Education, 1996a), and the annual report of children and youth with disabilities receiving special education (Oregon Department of Education, 1996b). The population estimates are reported in age ranges, and we used these to determine the percentage of the population reported on the average daily membership (ADM) and the percentage of the population reported on the annual special education count. The ADM is reported by grade level rather than age, and we estimated age for the ADM based on the idea that children starting kindergarten at age 5 years will begin their senior year of high school when they are age 17 years. Under the Individuals with Disabilities Education Act (IDEA), all children and youth with disabilities and requiring special education are eligible to receive it between birth and 21 years. There are, therefore, children in the Special Education Child Count (SECC) who are not included in the ADM (see Table 1).

In the context of state school funding, special education students are considered as a proportion of all school children, and the ADM and child count figures are used to calculate these percentages. We compared these numbers at each grade/age level to help establish proportions

Table 1. Proportion of Oregon Population (7/1/96) in Average Daily Membership (ADM) and Special Education Child Count (SECC)

Age Range	Population Estimate	Grade Level	ADM Count	Percent of Population	SECC** Count	Percent of Population	SECC Age Range
5-9	227,533	K*-4	204,411	89.84%	21,903	9.63%	5-9
10-14	223,118	5-9	210,214	94.22%	26,647	11.94%	10-14
15-17	134,209	10-12	104,523	77.88%	10,255	7.64%	15-17
Total 5-17	**584,860**		**519,148**	**88.76%**	**58,805**	**10.05%**	

* Kindergarten = actual count, each child counts as one, rather than 0.5 as in other counts

** Report of Children and Youth with Disabilities Receiving Special Education, Revised April 12, 1996.

of special education students that we might expect to find taking the test at 3rd, 5th, 8th, and 10th grades if all students participated in the assessment. These are the grades in which Oregon administers its assessments.

There are three separate data collections that were relevant:

- Average daily membership (ADM) collected and calculated by the Oregon Department of Education's (ODE) Office of School Finance.
- Oregon Statewide Assessment Program (OSAP) administered and managed by the ODE's Office of Assessment and Evaluation.
- The annual December 1, Special Education Child Count (SECC) conducted by the ODE's Office of Special Education.

The State Education Agency only collects data that are required by State or Federal mandate (see Table 2).

Each set of data is collected for a different purpose: ADM for school funding, OSAP to measure achievement, and SECC to distribute federal IDEA funding. Each collection represents a different point of view about what is meant by *all* students. ADM is reported annually for all children, kindergarten through 12th grade, enrolled in public schools. This information is reported by school districts in the form of counts by school and by grade. The OSAP is administered to all students in 3rd, 5th, 8th, and 10th grades, except those exempted from testing. The testing data file includes all answer sheets returned and is typically reported by grade, school, and district. The SECC collects information on each individual between birth and 21 years with an Individualized Education Plan (IEP) and receiving special education on December 1. This count includes not only children enrolled in public schools but also children in state schools for the deaf and blind, juvenile correction facilities, private agency programs, early intervention and early childhood programs (birth through 4 years), and home and parochial schooled children eligible for and receiving special education. Only the testing and special education data are available at the individual student level.

We selected the two extant data sets from assessment and special education for the 1995-96 school year. The reading and mathematics tests were administered during the spring. The number of students taking the tests includes all returned student answer sheets including those marked *modified* or *exempt*. Unmarked sheets are considered standard. Figure 1 displays the nonparallel nature of the separate data collections framed in the context of the population estimates. Our efforts to determine the figures used in the numerator and the denominator for calculating anticipated rates of test participation for all students including special education students brought the concern over neglected numerators and drifting denominators into sharper focus (Erickson

Table 2. Proportion of Special Education Child Count (SECC) in Average Daily Membership

ADM Report Grade Level	ADM Count	SECC*** Count	Percent	SECC Age in Years
K*	39,574.0	2,159	5.5%	5
1	42,038.9	3,039	7.2%	6
2	40,662.7	4,349	10.7%	7
3	40,644.0	5,665	13.9%	8
4	41,491.7	6,367	15.3%	9
5	41,922.3	6,071	14.5%	10
6	41,405.9	5,470	13.2%	11
7	41,927.4	4,907	11.7%	12
8	41,441.4	4,493	10.8%	13
Uncassified Elem.**	2,461.3			
9	41,056.9	4,082	9.9%	14
10	37,461.4	3,809	10.2%	15
11	34,268.2	3,106	9.1%	16
12	30,856.3	2,274	7.4%	17
Unclassified Secon.**	1,936.6			
Total	**519,149.0**	**55,791**	**10.7%**	

* Kindergarten = actual count of children, each child counts as one, rather than 0.5 as in other counts.

** Unclassified Elementary & Unclassified Secondary

*** Report of Children and Youth Receiving Special Education, Revised April 12, 1996.

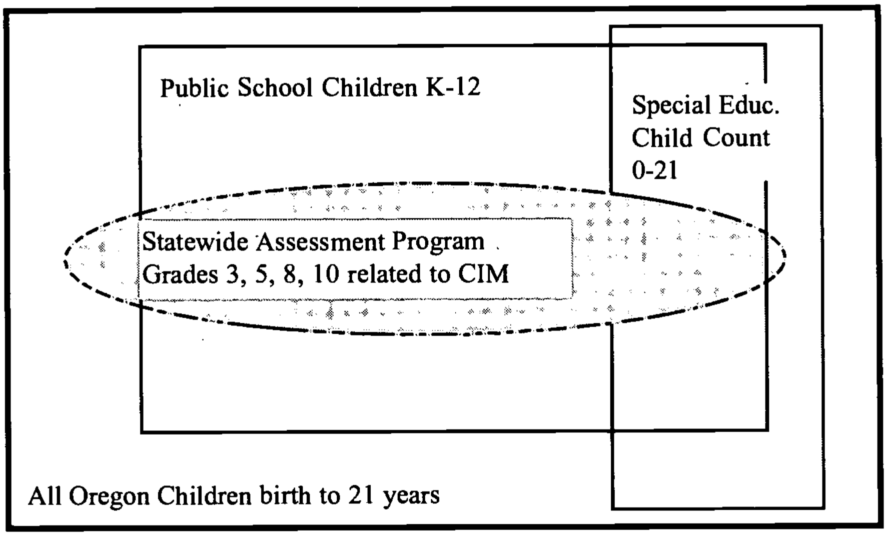
et al., 1996). Special education serves children who do not attend public schools and it is not clear whether they should be tested (see Figure 1).

Joining Two Extant Data Sets

The process of combining two extant computer databases would normally be very easy—sort the cases and merge by a common key (a specific field used for this purpose). The problem is that the two files do not share a common key. Various historical and legal conditions may obviate the development of common keys or identification numbers; likely the lack of common keys can be explained simply as the result of the files being created in two different offices within a state educational agency — assessment and special education.

There are nearly 160,000 student assessment records produced annually in Oregon. These records

Figure 1. Testing in Oregon Addresses Public Schooled 3rd, 5th, 8th, and 10th Graders


Public School Children K-12
Special Educ. Child Count 0-21
Statewide Assessment Program Grades 3, 5, 8, 10 related to CIM
All Oregon Children birth to 21 years

are divided somewhat evenly among grades 3, 5, 8, and 10, the grades tested on the statewide assessment. There were 35,000-40,000 records per grade level in the testing files and 60,000-65,000 records in the special education file. Oregon calculates basic school support for special education students using a weighted formula with a cap set at 11% of the state's average daily membership overall. The incidence of students with IEPs is actually higher between 3rd and 5th grade: 13.9% at 3rd grade, 14.5% at 5th grade, 10.8% at 8th grade, and 10.2% at 10th grade. We concluded that if all students participated in the assessment, we would find these proportions in the testing database.

Our basic formula for determining the proportion of special education students taking the test was suggested by Erickson et al. (1996): "the number of students with disabilities who take the test, divided by the population of all students with disabilities at the particular age or grade level being tested" (pp. 4–5).

In this study we joined individual student data from the OSAP and the SECC. We limited our investigation to fifth and eighth grade testing of reading and mathematics. Each of the two types of databases followed separate and idiosyncratic rules for the formation of both records and keys. As such it was impossible to merge the two outright. Steps first had to be taken to form a common key with which to blend the two files.

Keys are particular fields that allow for the sorting and matching of data contained in separate records. A number, e.g., social security number, is the preferred type of field for use as a key, but any alphabetic or numeric combination of characters may be used.

The merge key must be common to both files in order for the match-merge procedure to function. Without overlapping keys, the files cannot be match-merged. Listed below are the common fields used in the formation of merge keys for each of the two file types:

1. Special education database: This file uses a full last name, first name, and middle initial for each individual.

2. Test databases: Only 11 characters of last name and 7 characters of the first name are available.

In the Oregon files, the situation becomes more complex because the greatest common key form is limited by the test database, ergo a sub-string of the full special education name must be used. The special education last name (11 characters) is concatenated with the first name (7 characters) to form the fully-qualified match-merge key field. The resultant key is 18 characters, all alphabetic. All sorting and merging operations discussed below are based on this 18-character key.

One would think that there would be considerable overlap between individuals using this key (last 11 plus first 7), that is, two students would appear to match with only a partial first and last name. Ironically, only several identical keys resulted from using this concatenated match-merge key. The presence of these identical keys produced a system warning, but not an error. So we continued.

Standard merge procedures typically take a sorted list in the first or primary file and then look in the (previously sorted) second or secondary file for a corresponding match. If keys in the two files match, then a merge of two records is made. This generic merge model is somewhat different here in that both files are treated as equal—there is not a primary and lesser or secondary file. This tactic was used because both files were deemed equally important. We did not want to err on the side of one or the other in terms of including or not including a given subject's record.

Exact matches in the merge procedure represent a hit; all characters in both keys are the same. Not all merge instances result in a hit. There are near misses and complete misses. The computer algorithm that we used attempted to resolve these, but in the end, only exact complete matches successfully exited the match-merge procedure. The bottom-line: Only exact matches had both test score data and special education demographic data.

Given that we had to start with extant databases that were not merge compatible, we made the best of a bad situation. This approach was more brute force than anything, and we relied on the power of a large mainframe computer to form a solution. Our examination involved the following steps or phases. First, we determined the amount of overlap that could be identified between the two files. Next we examined what we might say about the matched or "in-both" records and the non-matched or "testing only" records related to the statewide reading and mathematics assessments. Then we determined the amount of confidence that we were willing to place in the identified overlap. We examined sources of inaccuracy in matches that achieved questionable confidence. Finally, we examined the performance of subgroups of students taking the test.

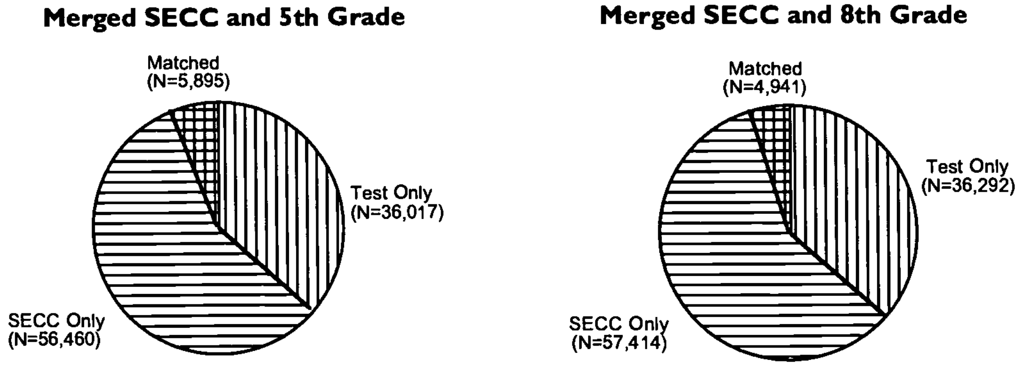
Findings

Overlap

In our first phase, we determined the amount of overlap that could be identified between the two files, special education and testing. Our initial action was to merge the special education data with the fifth grade reading and mathematics test data file and then the special education data with the eighth grade reading and mathematics test data file. We employed the match and merge procedure described above, using the 18-character name key common to both testing and special education files. We obtained two merged files including both special education and testing data. The results of these two merge procedures are portrayed in Figure 2. Each merged file contained all of the records from each component file. The fifth grade file contained 41,912 records, the eighth grade file contained 41,233 records, and the total special education file contained 62,355 records (see Figure 2).

We uncovered several of the problems anticipated by Erickson et al. (1996). The SECC is a snapshot report taken annually on December 1. The statewide reading and mathematics assessments are given the following April. In Oregon, the State Education Agency (SEA) has no unifying student record system containing student identification (ID) code to match files and no primary source file to use in verifying matches. The testing data set includes grade level because the tests are administered at grades 3, 5, 8, and 10. The SECC does not include a grade

Figure 2. Name Match and Merge of Extant Data Sources, Testing Data, and Special Education Child Count Data for the 1995-96 School Year



field, relying instead on birth dates and age calculations. Special education students may not always have a grade level designation because of the ungraded nature of some special education programs.

The testing data provided figures on the proportion of special education students taking the test. On the assessment answer sheet one section provides a place to indicate any educational programs in which the student participates. Teachers may assist students in grades 3 and 5 to mark their answer sheets and rely on 8th and 10th graders to mark independently. Programs include Title 1 reading, Title 1 math, migrant education, English as a second language, and special education. There are two separate answer sheets at fifth and eighth grades: one for reading and one for mathematics. Both have a section to mark program participation. The testing contractor combines reading and mathematics data into a common file for fifth grade and another for eighth grade. These files contain a single set of program codes that merges information from the two forms when the student takes both tests. The merged file program data indicated that 4,615 fifth graders and 3,051 eighth graders were in special education. Part of our task was to attempt to verify this information.

Following the formula suggested by Erickson et al. (1996) and working backwards, we obtained the numerator and denominator needed to execute it. We employed the formula with Oregon's fifth and eighth grade reading and mathematics assessments through the following steps. First, we identified the population of all students with disabilities who were age 10 for the grade 5 tests and age 13 for the grade 8 tests. The age-to-grade calculation was based on children who reach age 5 years on or before September 1 and enter kindergarten, adding one year and one grade level for each year in school. We calculated age using the birth year, month, and day from the special education file subtracted from September 1, 1995. SPSS Base 7.5 for Windows provides a function called YRMODA which conducts calculations using date fields for year, month, and day (1997). The files were merged on a mainframe computer. Some analysis was then done on the mainframe and on a personal computer with SPSS for Windows.

We then selected special education students from the file who were age 10 for the fifth grade test (N = 6,071) and age 13 for the eighth grade test (N = 4,493). During the merge procedure a descriptive marker variable was created that identified merged records containing *data* from the testing file with *Y* for *yes* and *N* for *no*. With these marker variables we were able to identify the number of students with disabilities who took the test for the participation rate formula. For the *Y*s we also checked the *program field* from the merged testing file to learn how many students from the special education file had the special education bubble marked on a testing answer sheet. Table 3 shows that 47.5 % of age 10 special education students were in the fifth grade testing file, and 38.6% of age 13 special education students were in the eighth grade testing file. When special education students took the test they did not always have special education marked as the program field. Only 63.1% of the fifth grade matched records had

Table 3. Selected Special Education Students at Age/Grade Level Found in Testing File

	Age/Grade	In Testing File		Test Program Field
	SE Count	No	Yes	Coded SE
5th Grade (10 Years)	6,071	3185	2886	1,820
Percent		52.5%	47.5%	63.1%
8th Grade (13 Years)	4,493	2,758	1,735	916
Percent		61.4%	38.6%	52.8%

special education marked in the program field and 52.8% of the eighth grade matched records had it marked (see Table 3).

One potential problem with the approach of selecting special education students at the grade-age tested is that some of these students may have repeated a grade or started first grade late. It is possible that some special education and general education students were actually 11 years old when they took the test. These students would not be represented in the percentages of special education students taking the test. Because of this possibility, we decided to analyze the merged files in more detail.

Characteristics of Students in Matched and Test Only Groups

In phase two, we examined what we might say about the matched or in-both records (i.e., students with both special education and test files) and the non-matched (special education only) or testing only records related to the statewide reading and mathematics assessments. The system files that blended special education data with fifth and eighth grade testing data used a minimal alphabetic merge key that was the concatenation of the first 11 characters of the last name and the first 7 characters of first name. Additionally, special variables were created to track (a) membership in the special education file, (b) presence in the fifth and eighth grade testing files respectively, and (c) presence in both files, the matched records. Records in the system file that were flagged as in-both represented students who both appeared in the special education file and took the reading and/or mathematics assessment.

We identified 5,894 matched records for fifth grade or 14% of the testing records (N = 41,912) and 4,941 matched records for eighth grade or 12% of the testing records (N = 41,233). If all students, including all special education students, took the tests, we expected to find approximately 6,000 special education students taking the fifth grade test and 4,500 special education students taking the eighth grade test. These numbers correspond to 14.5% and 10.8% of the average daily membership (refer to Table 2). The name-matched records came surprisingly close to the expected figures if all students, including special education students, participated.

The matched records provided an operational set of records that were found in both the special education file and the grade level testing files (fifth or eighth grade). We referred to these records as in-both, meaning found in both data sources. The balance of the testing records, those without matching special education records were called testing only, meaning the students had only a testing record from the testing file.

We then examined the two groups, in-both and testing-only, on three student characteristics reported on the testing bubble sheets: (a) special education membership indicated in the program fields, (b) student age calculated from the date of birth in the testing file and for in-both records age calculated from to the date of birth in the special education file, and (c) exclusion (modified and exempt) indicated as conditions of testing. Table 4 shows that 46.3% of the in-both fifth grade records had the special education program marked, and 34.6% of in-both eighth grade records had special education marked. In the testing files 4,615 fifth grade records and 3,051 eighth grade records had special education indicated in the program fields. Only 59.2% of fifth grade testing records and 56.0% of eighth grade testing records that had special education marked in the program field were located in the in-both group. Special education membership did not fully agree between the two separate data collections. This lack of agreement raised

Table 4. Program Code with In-Special Education Child Count

		5th Grade Testing File				
		In-Both		**Test Only**		**Total File**
Test Coded SE	Yes	2,730	46.3%	1,885	5.2%	4,615
		59.2%		40.8%		
	No	3,164	53.7%	34,133	94.8%	37,297
		8.5%		91.5%		
	Total	**5,895**		**36,017**		**41,912**
		14.1%		**85.9%**		
		8th Grade Testing File				
		In-Both		**Test Only**		**Total File**
Test Coded SE	Yes	1,709	34.6%	1,342	3.7%	3,051
		56.0%		44.0%		
	No	3,232	65.4%	34,950	96.3%	38,182
		8.5%		91.5%		
	Total	**4,941**		**36,292**		**41,233**
		12.0%		**88.0%**		

questions about the accuracy of the program field in the testing file and the accuracy of the in-both membership (see Table 4). It seemed unlikely that a large number of students had changed their special education status between December (the special education count) and March (the reading and mathematics test administration window).

What about the age of students taking the fifth and eighth grade tests in the two groups? Table 5 shows counts of students by age calculated from the date of birth in the testing file. Student records from the in-both group counts are reported by a second age, calculated on the date of birth in the special education file. The birth date field in the testing file for the testing only group had missing data in 15% of the fifth grade records and 13.4% of the eighth grade records.

Table 5. Ages for Testing Groups: In-Both and Test Only

5th Grade	**Test Only**		**In-Both**		**SECC**		
	Count	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Age**
birth to 7 years	140	0.4%	21	0.4%	341	5.8%	birth–7 years
8 years	35	0.0%	8	0.1%	178	3.0%	8 years
9 years	1,157	3.2%	141	2.4%	275	4.7%	9 years
10 years	26,130	72.5%	4,331	73.5%	2,886	49.0%	10 years
11 years	2,823	7.8%	1,004	17.0%	1,018	17.3%	11 years
12 to 21 years	77	0.2%	34	0.6%	1,196	20.3%	12–21 years
Invalid	98	0.3%	14	0.2%	0	0.0%	Invalid
Missing	5,558	15.4%	341	5.8%	0	0.0%	Missing
Total	**36,018**	**100.0%**	**5,894**	**100.0%**	**5,894**	**100.0%**	
Grand Total	**41,912**						

8th Grade	**Test Only**		**In-Both**		**SECC**		
	Count	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Age**
birth to 10 years	318	0.9%	38	0.8%	982	19.9%	birth–10 years
11 years	20	0.0%	0	0.0%	207	4.2%	11 years
12 years	1,188	3.3%	103	2.1%	231	4.7%	12 years
13 years	25,460	70.2%	3,245	65.7%	1,735	35.1%	13 years
14 years	4,193	11.6%	1,252	25.3%	1,163	23.5%	14 years
15 years	108	0.3%	45	0.9%	241	4.9%	15 years
16-21 years	7	0.0%	5	0.1%	382	7.7%	16–21 years
Invalid	122	0.3%	11	0.2%	0	0.0%	Invalid
Missing	4,876	13.4%	242	4.9%	0	0.0%	Missing
Total	**36,292**	**100.0%**	**4,941**	**100.0%**	**4,941**	**100.0%**	
Grand Total	**41,912**						

Since the birth date bubbles are marked by the student, missing data result when the student fails to respond to an item. Most of the students taking the test were at the expected age for the grade level tested. In the in-both group 17% of the students were 11 years old when they took the fifth grade test and 25% of the students were 14 years old when they took the eighth grade test. This percentage is slightly higher than that in the testing only group. The lack of agreement between the counts by age calculated from the testing file and counts by age calculated from the special education file reinforced concerns about the accuracy of the in-both membership.

Exclusion codes refer to coded responses on the testing answer sheets that indicate that the test was taken under nonstandard conditions. There were seven categories of exclusion in the 1996 test administration: absent, exempt Limited English Proficiency (LEP), modified LEP, exempt special education, modified special education, other (usually parent refused), and no exclusion. Table 6 displays the counts for the exclusion codes for both reading and mathematics and both grade levels tested. A larger proportion of students in the in-both group had special education *exempt* or *modified* marked. This was expected. The proportion of in-both records that had no exclusions was surprising. It seemed unlikely that 75-85% of students in the in-both group took the assessment without modifications or exemptions.

Program coding for special education lacked satisfactory agreement with special education membership from the child count data, and agreement between ages calculated on the testing birth date and the special education birth date was poor. The proportion of in-both, matched records that appeared in the testing file with no exclusion codes seemed suspicious. We decided to conduct a more thorough analysis of the quality of the matched records.

Confidence

In phase three, we evaluated the quality of matched records; we wanted to know whether they were true matches. We determined the amount of confidence that we were willing to place in the identified overlap by examining sources of error or inaccuracy. We first created a string made up of the first three characters of the last name and then (more or less randomly) chose a value of CAS for this string, indicating that the last name in selected records began with CAS. In this way we selected a range of student records for further scrutiny in terms of match versus non-match. Finally, we sorted and split cases by the source variables and listed the cases.

In order to better understand the problem with matching on the name field, we listed the name from the testing file and the name from the special education file for all 255 sampled CAS records. We examined the selected sample from the eighth grade testing file and found 255 records with a last name beginning with CAS (or 0.26% of combined file, N = 94,862). Hand matching from the printout gleaned an additional five cases. Figure 3 lists names as they appeared

Table 6. In-Both and Test Only Counts by the Various Exclusion Indicators

Reading Test 5th Grade

	Test Only	Percent	In-Both	Percent	Total
Absent	440	1.2%	104	1.8%	544
Exempt LEP *	348	1.0%	27	0.5%	375
Modified LEP	174	0.5%	16	0.3%	190
Exempt SE *	352	1.0%	632	10.7%	984
Modified SE	308	0.9%	600	10.2%	908
Other	142	0.4%	31	0.5%	173
No Exclusion	34,254	95.1%	4,484	76.1%	38,738
Total	**36,018**		**5,894**		**41,912**

Math Test 5th Grade

	Test Only	Percent	In-Both	Percent	Total
Absent	386	1.1%	83	1.4%	469
Exempt LEP *	209	0.6%	17	0.3%	226
Modified LEP	209	0.6%	20	0.3%	229
Exempt SE *	260	0.7%	494	8.4%	754
Modified SE	457	1.3%	681	11.6%	1,138
Other	141	0.4%	26	0.4%	167
No Exclusion	34,356	95.4%	4,573	77.6%	38,929
Total	**36,018**		**5,894**		**41,912**

Reading Test 8th Grade

	Test Only	Percent	In-Both	Percent	Total
Absent	763	2.1%	166	3.4%	929
Exempt LEP *	204	0.6%	17	0.3%	221
Modified LEP	121	0.3%	9	0.2%	130
Exempt SE *	189	0.5%	319	6.5%	508
Modified SE	175	0.5%	317	6.4%	492
Other	180	0.5%	33	0.7%	213
No Exclusion	34,660	95.5%	4,080	82.6%	38,740
Total	**36,292**		**4,941**		**41,233**

Mathematics Test 8th Grade

	Test Only	Percent	In-Both	Percent	Total
Absent	892	2.5%	171	3.5%	1,063
Exempt LEP *	138	0.4%	12	0.2%	150
Modified LEP	102	0.3%	8	0.2%	110
Exempt SE *	171	0.5%	294	6.0%	465
Modified SE	183	0.5%	255	5.2%	438
Other	197	0.5%	28	0.6%	225
No Exclusion	34,609	95.4%	4,173	84.5%	38,782
Total	**36,292**		**4,941**		**41,233**

* LEP = Limited English Proficiency, SE = Special Education

Figure 3. Sample Records in a Comparison of Names in Two Data Files (N = 255)

Source File	Name
8-Test Name:	CASSLE, MANUEL
SPED Name:	CASSLE, MANWELL
8-Test Name:	CASTENADA, CRISTA
SPED Name:	CASTENEDA, CRYSTAL
8-Test Name:	CASTILLEJA, RAYMOND
SPED Name:	CASTILLEJA, R.J.
8-Test Name:	CASTLE, TONY
SPED Name:	CASTLE, ANTHONY
8-Test Name:	CASTRO, MCKENZI
SPED Name:	CASTRO, C. MCKENZI

in the eighth grade testing file and the special education file (actual names have been modified in this example).

Names contained in the two files appeared to match based on a visual scan of the listing but failed the match based on the 18-character name key. The failure was due to minor changes in spelling, nicknames, and the use of initials in place of first or middle names. When the visual scan indicated a match and the name key did not, the record was coded a non-match. We restricted our work to electronic matching and merging during this study. We did not review individual records and enter corrections into the combined data file by hand. In this way we may have failed to find matches that were actually present. The basic conclusion from this initial exercise was that a longer name key would not drastically improve the match rate.

We examined the gender, birth date, and district fields to learn how information from these fields might inform a process of reconciling matches. We investigated further to determine the nature of the disagreement when one or more of the confidence fields did not match. Figure 4 shows several records with an 18-character name key match that contain non-matched corresponding fields. Fields in the testing file that did not agree with fields in the special education child count file appear in bold italics. (Again, names are altered in this example to maintain confidentiality.)

We established a confidence test for each record matched on the 18-character name key. We calculated a special confidence field that flagged the degree of agreement between the two

Figure 4. Examples of Records Without All Three Confidence Fields Matching

				BIRTH DATE							
NAME		GENDER		SE* CHILD COUNT			8TH GRADE TEST			DISTRICT	
Last	First	SE*	G8*	YY	MM	DD	YY	MM	DD	SE*	G8*
Backer	Jason	M	M	81	10	21	***80***	***10***	***12***	02059	***34023***
Beller	Steve	M	M	89	12	2	***81***	***5***	***16***	26001	***22129***
Blackbird	Sandra	F	F	89	2	10	***81***	***9***	***16***	20004	***24024***
Carter	Jesse	M	M	81	5	31	***80***	***12***	***12***	20052	***37001***
Davis	Christian	M	M	82	2	5	***81***	***7***	***13***	26001	***24024***
Garrison	Richard	M	M	90	12	19	***82***	***5***	***12***	04001	***26007***
Peters	Adriana	F	F	90	10	26	***96***	***10***	***22***	20004	20004
Sanchez	Manuel	M	M	81	3	16	.	.	.	10019	10019

SE = Special Education Count, G8 = Grade 8 Testing

matched records on three corresponding fields: gender, birth date, and district codes. As shown in Figure 5, approximately half of the in-both or matched records also had agreement between the gender, birth date, and district codes. This was done to provide further evidence that the matched records were true matches.

We began to view matches that did not agree on all three confidence fields as questionable matches. Without confidence that the special education status and the test score were from the same student, it would be unreasonable to evaluate special education student performance. We

Figure 5. In-Both Reliability Based on Name Key Matches Displaying Proportion of Complete Matches

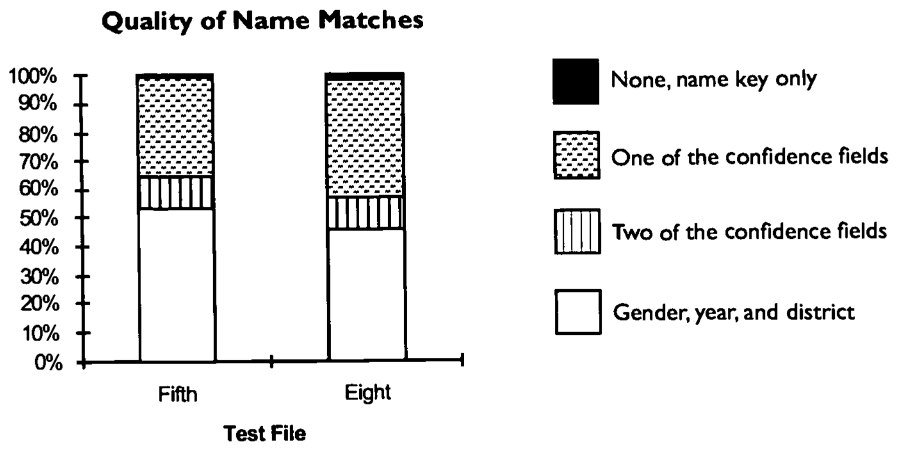


Table 7. Confidence for Matched Records between SE Child Count, 5th and 8th Grade Test Files

	In-Both 5th Grade		In-Both 8th Grade		
Degree of Match	**Count**	**Percent**	**Count**	**Percent**	**Type of Match**
Gender, year, and district	3,135	53.2%	2,293	46.4%	Good Match
Two of the confidence fields	663	11.2%	521	10.5%	Questionable
One of the confidence fields	2,042	34.6%	2,075	42.0%	Questionable
None, name key only	54	0.9%	52	1.1%	Non-Match
Total—Name Matches	**5,894**	**100.0%**	**4,941**	**100.0%**	

were reluctant to count records as true matches when the birth dates did not match. The counts and percentages for each confidence level are provided in Table 7. We developed the following three classifications for matches and designated each record in the system file as falling into one of the following categories:

- Good Matches: Records that matched on the 18-character name key and also matched on gender, birth date, and district were counted as good matches.

- Questionable Matches: Records that matched on the 18-character name key but matched on only one or two of the confidence fields were considered questionable matches.

- Non-Matches: Records appearing in the eighth grade testing file that matched on the name key only and had no matches on the confidence fields or did not match using the 18-character name key were viewed as non-matches.

This analysis brought the number of good matches, those about which we were confident, to 3,135 good matches in the fifth grade and 2,293 good matches in the eighth grade testing files. We could confidently place these numbers in the numerator to determine the number of special education students taking the state tests. Using the same grade-based age groupings as the basis for the denominator, that is, 6,071 for fifth grade and 4,493 for eighth grade, we arrived at much more conservative proportions (see Table 7).

Based on the good matches, 51.6% of special education students took the fifth grade test or returned an answer sheet and 51.0% of special education students took the eighth grade test. Three estimates of special education participation in statewide testing were calculated for this study, one produced by selecting special education students age 10 and 13 using the name key and locating matches with testing records, another by looking at all testing records and considering those with special education indicated in the program code fields, and finally one found by

Table 8. Three Estimates of Test Participation

Special	5th Grade		8th Grade	
Education	10 Years	Percent	13 Years	Percent
Grade based age	6,071		4,493	
Test Coded SE	4,615	76.02%	3,051	67.91%
In SE, also in test	2,886	47.54%	1,735	38.62%
Good Matches	3,153	51.94%	2,293	51.03%

calculating the proportion based on good matches. Table 8 shows all three approaches for comparison.

For the remainder of the study we conducted our analyses using the three confidence categories. Table 9 displays counts for the fifth and eighth grade tests for several descriptive fields from the test record: special education indicated in the program field, language fluency marked fluent, any exclusion marked in reading, any exclusion marked in mathematics, reading attempted

Table 9. Demographics for Confidence Groups

		Good Match		Questionable		Test Only		
		Count	Percent	Count	Percent	Count	Percent	
5th Grade								
Total in Group		3,135		2,705		36,072		41,912
Coded In SE	Yes	2,143	68.4%	583	21.6%	1,889	5.2%	
Language Fluency	Yes	2,736	87.3%	2,367	87.5%	30,592	84.8%	
Reading Exclusion	Yes	1,031	32.9%	374	13.8%	1,769	4.9%	
Math Exlcusion	Yes	971	31.0%	345	12.8%	1,667	4.6%	
Attempted Rdg	Yes	2,686	85.7%	2,506	92.6%	31,134	86.3%	
Attempted Math	Yes	2,792	89.1%	2,553	94.4%	31,283	86.7%	
8th Grade								
Total in Group		2,293		2,597		36,343		41,233
Coded In SE	Yes	1,359	59.3%	345	13.3%	1,347	3.7%	
Language Fluency	Yes	1,982	86.4%	2,292	88.3%	31,784	87.5%	
Reading Exclusion	Yes	602	26.3%	254	9.8%	1,637	4.5%	
Math Exlcusion	Yes	530	23.1%	230	8.9%	1,691	4.7%	
Attempted Rdg	Yes	2,011	87.7%	2,686	103.4%	31,081	85.5%	
Attempted Math	Yes	2,017	88.0%	2,792	107.5%	31,232	85.9%	

marked *yes*, and math attempted marked *yes*. Attemptedness is a quality indicator. The testing contractor responsible for scanning the testing answer sheets and creating an electronic test file on magnetic tape produced the attemptedness. The student must have marked valid responses to at least five test items to be coded *yes*, indicating that he or she attempted the test. Program coded special education, reading exclusions, and math exclusions all occurred more frequently in the good match group. Gender also showed differences between the categories. Table 10 shows that a higher proportion of good match records were males (64.6% in fifth grade and 64.8% in eighth grade). The ratio of males and females in the good match group agreed with gender breakdowns consistently reported for special education students.

Test Performance

Ultimately, we examined test performance for the three confidence categories of students taking the fifth and eighth grade reading and mathematics assessments. Oregon is preparing, as are many states, to report assessment results disaggregated by groups of special interest. The newest Title 1 requirements specify reporting the performance of subcategories of students. There is great interest in learning how special education students fare on the new high academic standards

Table 10. Gender by Testing Only, Questionable, and Good Matches

5th Grade Test		Testing Only	Questionable Count	Good Match Count	Total Count
Gender	Female	18,352	820	1,111	20,283
		50.9%	30.3%	35.4%	48.4%
	Male	17,624	1,875	2,024	21,523
		48.9%	69.3%	64.6%	51.4%
	Missing	96	10	0	106
		0.3%	0.4%	0.0%	0.3%
	TOTAL	**36,072**	**2,705**	**3,135**	**41,912**

8th Grade Test		Testing Only	Questionable Count	Good Match Count	Total Count
Gender	Female	18,232	796	807	19,835
		50.2%	30.7%	35.2%	48.1%
	Male	17,888	1,763	1,486	21,137
		49.2%	67.9%	64.8%	51.3%
	Missing	223	38	0	261
		0.6%	1.5%	0.0%	0.6%
	TOTAL	**36,343**	**2,597**	**2,293**	**41,233**

that have been adopted by the State Board of Education. The stakes are high and reports need to be accurate and straightforward, not misleading. The staff psychometrician for the OSAP recommended a standard requiring a valid response for 75% of the items in the test for analyzing testing performance for this study (personal communication March, 1997). Using the test attemptedness of five valid responses would allow scores to be included in the analyses that had an unacceptable standard error of measurement. Oregon's assessment assumes that each student responds to all of the items. The test is referred to as a power test not a timed test, and extended time is allowed for all students making progress on the test. We followed the recommendation and included only testing records that met the 75% criteria to be included in the analyses. Essentially, valid responses were either correct or incorrect responses. There were two exceptions. In some cases, a student marked two bubbles for the same item and an asterisk appeared in the field for that item, and in another case, an item was eliminated during the scoring process and the field contained an *X*. All of these cases were considered valid because the student responded to the item. Figure 6 shows the proportion of valid item responses for in-both and test only groups. More than 80% of students had valid responses for at least 75% of the items for both groups in both subjects (reading and mathematics) and at both grade levels. Table 11 provides a detailed breakdown for the three confidence categories across three levels of valid responding.

Next we examined the performance of special education students on the fifth and eighth grade statewide reading and mathematics assessments. We examined performance on the test for all three confidence groups at both fifth and eighth grades. Table 12 provides assessment results for fifth grade reading and mathematics. Table 13 provides results for eighth grade reading and mathematics.

Figure 6. Proportion of Valid Item Responses For In-Both and Testing Only Groups

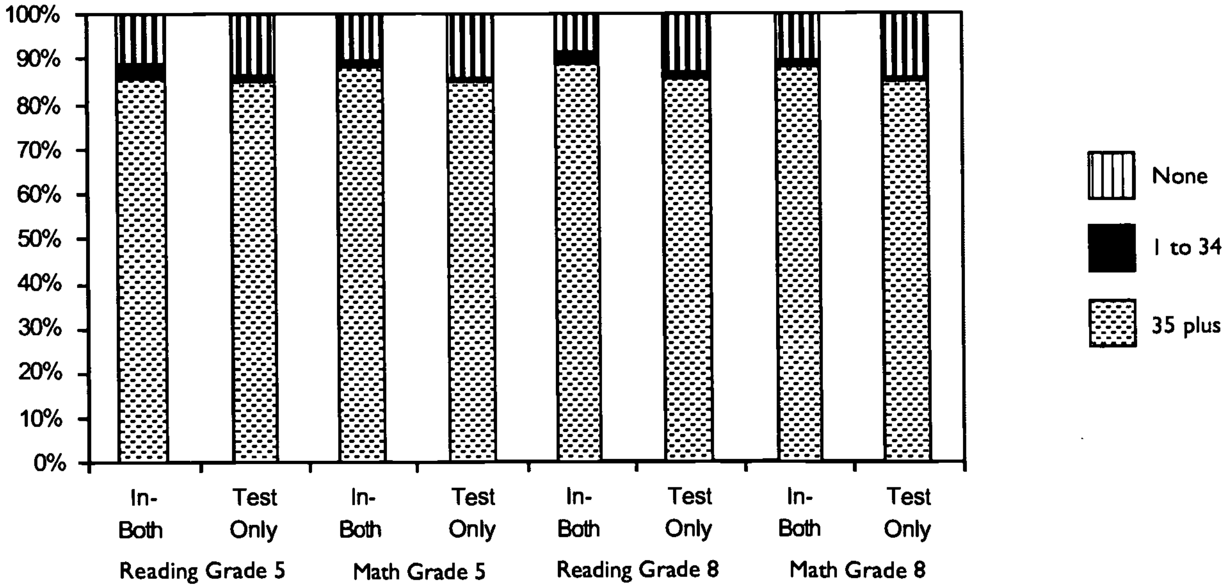


Table 11. Confidence for Matched Groups with Count of Valid Test Responses for Total Group

5th Grade Test		**Count of Valid Responses to Test Items**			
Reading	**Valid Responses**	**Testing Only**	**Questionable**	**Good Match**	**Total**
	Zero	4,925	193	440	5,558
	Less than 75%	432	64	158	654
	75% +	30,715	2,448	2,537	6,212
	Total	36,072	2,705	3,135	41,912
Mathematics	**Valid Responses**	**Testing Only**	**Questionable**	**Good Match**	**Total**
	Zero	4,767	146	332	5,245
	Less than 75%	441	60	128	629
	75% +	30,864	2,499	2,675	5,874
	Total	36,072	2,705	3,135	41,912
8th Grade Test					
Reading	**Valid Responses**	**Testing Only**	**Questionable**	**Good Match**	**Total**
	Zero	5,254	223	281	5,758
	Less than 75%	203	24	50	277
	75% +	30,886	2,350	1,962	6,035
	Total	36,343	2,597	2,293	41,233
Mathematics	**Valid Responses**	**Testing Only**	**Questionable**	**Good Match**	**Total**
	Zero	5,228	219	271	5,718
	Less than 75%	305	33	52	390
	75% +	30,810	2,345	1,970	6,108
	Total	36,343	2,597	2,293	41,233

The Oregon assessment results are reported as scale scores ranging between 150 and 280. The assessment design is based on Rausch Item Theory (RIT). Each year school, district, and state averages are reported for all assessments. Not all scores are included in the averages. There are two criteria for excluding scores from analysis and reporting:

- One of the six exclusion codes is flagged. Records marked *absent*, *modified* (LEP or SE), *exempt* (LEP or SE), or *other* are excluded from reporting.

- There is no marker in the Test Attemptedness field for the scores (separate flags for mathematics and reading) to be averaged.

These same criteria are used to exclude scores from the calibration process conducted on a

Table 12. Fifth Grade Assessment Results for 1995-96

		Reading				
		N	**Mean**	**SD**	**Minimum**	**Maximum**
Standard Inclusion	Testing Only	30,293	218.70	10.99	165	261
	Questionable Match	2,290	216.12	11.67	165	261
	Good Match	2,051	208.97	11.37	173	250
	Total	34,634				
75% No Exemptions	Testing Only	29,966	218.93	10.80	185	261
	Questionable Match	2,255	216.47	11.36	188	261
	Good Match	1,962	209.66	10.95	185	250
	Total	34,183				
75% Only	Testing Only	30,715	218.58	10.98	185	261
	Questionable Match	2,448	215.38	11.75	188	261
	Good Match	2,537	208.34	10.61	185	259
	Total	35,700				
75% SE Modified	Testing Only	283	204.82	8.31	188	235
	Questionable Match	129	202.82	8.47	189	238
	Good Match	430	204.04	7.49	188	230
	Total	842				
75% SE Exempt	Testing Only	142	203.27	8.68	188	230
	Questionable Match	41	200.88	7.47	189	222
	Good Match	126	202.69	7.17	189	223
	Total	309				
New Standard 9/19/96			215.00			
Prior "Proficient" Standard			204 to 222			

		Mathematics				
		N	**Mean**	**SD**	**Minimum**	**Maximum**
Standard Inclusion	Testing Only	30,410	214.88	9.66	158	267
	Questionable Match	2,326	213.08	10.07	177	267
	Good Match	2,125	207.54	9.36	177	254
	Total	34,861				
75% No Exemptions	Testing Only	30,088	215.06	9.51	187	267
	Questionable Match	2,294	213.33	9.87	188	267
	Good Match	2,061	207.92	9.16	189	254
	Total	34,443				
75% Only	Testing Only	30,864	214.80	9.62	185	267
	Questionable Match	2,499	212.47	10.11	188	267
	Good Match	2,675	206.98	8.92	189	256
	Total	36,038				
75% SE Modified	Testing Only	324	204.35	7.75	191	236
	Questionable Match	148	202.78	6.83	193	223
	Good Match	488	204.14	6.79	192	231
	Total	960				
75% SE Exempt	Testing Only	105	203.06	8.46	191	237
	Questionable Match	34	201.32	7.8	191	220
	Good Match	107	201.78	6.61	189	219
	Total	246				
New Standard 9/19/96			215.00			
Prior "Proficient" Standard			207 to 225			

Table 13. Eighth Grade Assessment Results 1995-96

Reading

		N	Mean	SD	Minimum	Maximum
Standard Inclusion	Testing Only	30,460	232.17	11.26	189	273
	Questionable Match	2,274	230.28	11.69	189	273
	Good Match	1,635	219.97	9.69	189	262
	Total	34,369				
75% No Exemptions	Testing Only	30,345	232.25	11.20	199	273
	Questionable Match	2,262	230.37	11.63	201	273
	Good Match	1,606	220.16	9.61	198	262
	Total	34,213				
75% Only	Testing Only	30,886	232.02	11.30	189	273
	Questionable Match	2,350	229.90	11.81	201	273
	Good Match	1,962	219.46	9.50	198	262
	Total	35,198				
75% SE Modified	Testing Only	160	217.59	8.50	203	250
	Questionable Match	44	216.11	8.03	207	240
	Good Match	268	216.46	7.88	201	242
	Total	472				
75% SE Exempt	Testing Only	53	214.98	9.01	189	247
	Questionable Match	17	213.88	10.59	204	238
	Good Match	67	216.46	9.7	201	247
	Total	137				
New Standard 9/19/96			231.00			
Prior "Proficient" Standard			216 to 234			

Mathematics

		N	Mean	SD	Minimum	Maximum
Standard Inclusion	Testing Only	30,443	231.31	9.54	184	283
	Questionable Match	2,292	230.22	9.82	203	207
	Good Match	1,693	222.05	7.79	204	263
	Total	34,428				
75% No Exemptions	Testing Only	30,264	231.41	9.45	184	283
	Questionable Match	2,278	230.30	9.79	206	270
	Good Match	1,675	222.17	7.72	205	263
	Total	34,217				
75% Only	Testing Only	30,810	231.24	9.51	184	283
	Questionable Match	2,345	230.00	9.86	206	270
	Good Match	1,970	221.61	7.56	205	263
	Total	35,125				
75% SE Modified	Testing Only	157	219.68	6.63	206	247
	Questionable Match	29	221.72	6.73	213	236
	Good Match	211	218.85	5.70	209	238
	Total	397				
75% SE Exempt	Testing Only	56	217.75	8.81	185	241
	Questionable Match	12	215.00	2.09	213	219
	Good Match	70	217.34	5.26	209	234
	Total	138				
New Standard 9/19/96			231.00			
Prior "Proficient" Standard			221 to 239			

sample of tests returned (S. Choi, personal communication, June 2, 1997). Test calibration is done prior to producing and reporting averages.

We selected four testing data subsets for analysis in addition to the standard group typically included in reporting averages. The standard inclusion involves all records with no exclusion codes and test attemptedness equal to one, indicating that the test was attempted. We then selected records with no exclusion codes and meeting the criteria of 75% valid item responses. This selection provided a comparison of the standard inclusion and the 75% criteria. Next, we selected testing records regardless of the presence of exclusion codes but still including the 75% criteria. We then selected two more sets for reporting, both included the 75% criteria. The reported scale score means for records marked exempt special education showed means ranging between 200.88 and 203.27 for fifth grade and 215.00 and 217.75 for eighth grade. Under the previous standard for proficiency the fifth grade scores fell below proficiency and only the eighth grade exempt special education reading mean for the good match group reached into proficient range grade (refer to Tables 12 and 13).

Prior to 1996, scores were judged as Basic, Proficient, and Advanced. On September 19, 1997, the Oregon State Board of Education adopted high academic standards and specified scale scores required to meet the standard at 3rd, 5th, 8th, and 10th grades. The criteria for proficiency and for meeting the standard are specified at the bottom of Tables 12 and 13 as a point of reference.

Making generalizations about the performance of students in the four selected groups would be suspect. Some interesting observations can, however, be made about the nature of indicators used for reporting. If these data are representative, it appears that lower scores may be associated with special education status. The good match confidence group scores displayed in Tables 12 and 13 are lower than those for the testing only group in the standard inclusion scale score means.

It is the special education exempt set that raises the most questions. For the fifth grade assessment, there were 309 students in reading and 246 students in math with valid responses to at least 75% of the items. For the eighth grade assessment there were 137 in reading and 138 in mathematics. This observation raises questions about the exclusion codes. In particular, on the fifth grade reading assessment 142 students were in the testing only group, met the 75% criteria, and were marked exempt special education. The circumstances surrounding these 142 cases are curious. These 142 scores were exempt from reporting summaries and in fact would not have been included in the sample selected to conduct the test calibration. Marking the tests exempt excluded them from analysis and reporting. One can only speculate about how the decision was made to exempt a test that contained 75% valid responses.

The program field's bubbles can be marked by the teacher in 3rd and 5th grade testing and can be filled in by students in the 8th and 10th grades. It is possible that there is some confusion about when to mark the special education program field. School averages are often reported in local papers and receive a lot of public attention. There has been a growing concern about the effect that special education students will have on their school's averages. It is generally thought that special education students will not do well on the tests and that they will bring down the school scores.

In discussing these findings with school principals, we learned that there are perceived incentives for indicating that a student took the test under modified conditions or was exempted from testing. School communities feel beat up by the testing reports that appear in their local media. They would often do anything to avoid having low test scores reported, even excluding students they believe will bring down the school averages. It is possible that the test file fields are filled in not solely on the basis of the student's status in special education or specifications of the IEP, but on the effect their score is expected to have on the school averages. We also learned in focus groups with teachers and school administrators that the decision to have special education students participate in testing is handled inconsistently. Sometimes the decision is made the week before testing by the classroom teacher and special education teacher based on their judgment of student readiness. Reschly (1993) uses the term "unwarranted exclusion" to refer to the "arranged non-participation in state or national assessment programs involving students for whom the assessment content is appropriate to curriculum goals pursued in their educational programs and the receptive or expressive language demands of the assessment tasks are within the student's behavioral repertoire" (p. 41).

The testing file contains missing data in the demographic and program fields that make it difficult to fully account for all of the student records when breaking the tested population into subgroups. Missing birth years, fields indicating special education status for students not appearing in the December count, student records marked exempt from testing with responses to items, and student records marked exempt but also attempting the test present problems with interpreting both the participation and performance for special education students.

Discussion

This investigation addressed three main areas of concern: (1) electronic matching of student records from two extant data files, and the quality of data contained in the demographic and program fields of the testing files, (2) the participation rate of students with disabilities in statewide assessment, and (3) testing and reporting on the performance of special education students.

We focused much attention on the problems inherent in electronically merging student records in extant data files. Implicit in this analysis is the need to clarify two issues: (1) which special education students actually took the test, and (2) which special education students were eligible to take the test but did not. Ultimately our goal was to implement the formula identified by NCEO for calculating the participation rate of special education students in statewide assessment programs (special education students tested/students eligible to take the test) and examine the performance of special education students on a statewide assessment.

We found that in Oregon, "all students" has different meanings in separate data collection efforts. For the special education child count, all students referred to children birth through 21 years of age with disabilities and requiring special education on December 1, and included children and students educated in preschools, corrections, state schools, private agencies, their homes, and parochial schools. For the reporting of average daily membership, all students referred to an average of students enrolled in public schools for kindergarten through 12th grade throughout the year. For statewide assessment, all students referred to the entire set of answers sheets returned from the spring administration of the reading and mathematics assessments. Separate data collections used separate definitions.

Electronic Matching

We started with two data sets: the Special Education Child Count (SECC), collected annually on December 1, and the Oregon Statewide Assessment Program testing data for Reading and Mathematics, administered in April, 1996. Early on, we established an expectation about how many special education students might be available based on a proportion of the student population or average daily membership enrolled in special education. Based on the proportion of students in the state's average daily membership (ADM) that is reported on the December 1 count, we anticipated identifying 6,000 special education students in fifth grade and 4,500 in eighth grade if all students were represented. We calculated three separate participation rates for each grade ranging between 38% and 76%. We were unable to verify any of the rates as the correct ones.

The electronic match-merge procedure was critical to identify the special education students who took the test. To establish the match we developed an 18-character name key with three confidence fields. While we were able to identify nearly 6,000 matches in the fifth grade file and nearly 5,000 matches in the eighth grade file, roughly half of these matches could be depended on as true matches. In other words, electronic matching between two files that do not contain a common unique identifier was laden with problems. A single pass with the existing algorithm was insufficient, yielding less than half of the expected matches.

In the absence of a student ID, the name key alone was not sufficient to produce true matches. Additional fields common to both files increased the confidence that could be placed in a given match. The accuracy of each field is a major factor in the quality of a match and merge between to extant files.

To increase the ability to identify matches in the files, we believe that we will need to add several steps to the electronic match-merge process. We believe that we could increase the matches with confidence through iterative matching with multiple algorithms and additional passes between the files. While it would be labor intensive, we suspect that adding a visual scan of similar records would pick up matches missed by the electronic algorithms. Finally, it may be necessary to conduct a phase of actual data cleanup by going back to districts for missing data and corrupt codes.

We analyzed several marker variables forming part of the data file in much the same way as Merwin (1993) reports on Jaeger's (1992) analysis of international studies presented for the American Educational Research Association, Division D. Jaeger found that the population variance of outcomes is heavily influenced by subgroups such as percentage of children living in single parent families, percentage of youth economically active, and percentage of children from single parent families living in poverty. Merwin's assertion is that these variables are likely to contribute more to the outcomes than the percentage of students with disabilities in statewide testing programs. In our case, much of the score variation related to either the special education good match group or the questionable match group. For the student records considered good matches, test scores were lower. We didn't read much into this finding, however, because we were unable to resolve the questionable matches.

In Oregon, data collections are operated by different offices under separate statutes and federal regulations. Data are maintained in separate unrelated files. There is a missing link, a unique student identifier that effectively relates the elements from the separate collections. These presented obstacles in the analysis of the data.

Special Education Membership

We found an unsettling level of disagreement between various indicators of special education membership. There was limited agreement about membership between the December count and the testing data file. We hypothesized that school communities wished to avoid having low test scores reported and worried about including students who, they believe, would bring down the school averages. It is possible that the test file fields are marked not solely based on the student's status in special education but on the effect the student's score is expected to have on the school averages.

The program fields can be marked by the teacher in 3rd and 5th grade testing and by the students in 8th and 10th grades. It is possible that there is some confusion about when to mark the special education program field. School averages are often reported in local papers and receive a lot of public attention. There has been a growing concern about the effect that special education students will have on their school's averages.

We also found disagreement between the birth date in the testing file and the corresponding birth date in the special education file. Both birth date and membership in special education are essential pieces of information for examining test participation and performance for special education students.

In the age-to-grade calculation, results indicated that 47.5% of fifth grade (age 10 years) and 38.6% of eighth grade (age 13 years) special education students returned testing answer sheets. The age-to-grade calculation may miss older special education students returning answer sheets. According to the fifth and eighth grade testing files 76.0% and 67.9% of special education students took the test. When good matches were considered, 51.9% (fifth grade) and 51.0% (eighth grade) of special education students took the test.

Test Performance

Scale scores for special education students reported in this study were consistently lower than the testing only group. A number of students had valid responses to at least 75% of the items and yet were marked exempt from testing. Questions about measurement error appeared in two areas: student demographic including program participation information and test performance. The demographic and program participation data contained inaccuracies and missing data that interfered with determining participation rates and student performance confidently. More importantly, we uncovered confusion about student records marked exempt from testing, the number of test items to which students responded, and scores assigned. This area of analysis warrants further review along both quantitative and qualitative avenues, once questionable matches are satisfactorily resolved. Student answer sheets marked exempt are eliminated from analysis and reporting. Specifically, they are excluded when records are sampled for test calibration and when summaries are reported at the school, district, and state level.

Ysseldyke, Thurlow, McGrew, and Vanderwood (1994) noted that assigning zeros to students who are excluded from testing was not found acceptable by many state department personnel. We would argue that zeros arise in many ways with a data file and that physical absence is only one way to corrupt the results. Consideration of student motivation and concurrent analysis of test attemptedness should be part of all statewide assessments and outcome reports.

Recommendations

Demographic and program participation data will be used to group scores and consider the effects of age, program services, socioeconomic status, and language proficiency. Confidence in these data is essential. More importantly, student progress toward criterion standards will increasingly influence school and program funding, student certification of mastery, and, in some states, teacher retention and salary incentives. The level of error that will be considered acceptable for these purposes will need to be clearly specified and consciously reported. We have learned that there are special education students with test scores. To understand their meaning, we will need to do a better job of identifying groups and we will need to learn more about the responses to items and the marking or test-taking mechanics.

Ysseldyke, Thurlow, McGrew, and Vanderwood (1994) noted that "there is a need to monitor exclusion of students with disabilities. Large scale assessments employ monitors to ensure that standardized procedures are followed" (p. 13). Such monitoring should occur also in the encoding of data, analysis of outcomes, and development of reports. Total tests should be validated with checks of students who complete all subtests (or items) versus those who complete only some of the subtests (or items) to ensure the sampling plan of those reported is intact and provides a comparability across subtests. In Merwin's (1993) analysis "intra-unit comparisons are used to show change in the aggregate index over time" (p. 30); we would add that such comparisons also are critical in understanding the school profiles (strengths and weaknesses). Our findings confirm several points emphasized in the careful and systematic analyses of the issue cited throughout their report.

The challenges of successfully matching and blending two extant files absorbed much of the effort in this study. While special education students may be physically present in the school and may even take the test, this presence does not necessarily constitute electronic presence in a manner that can be easily traced. Several recommendations come to mind as ways of mending this problem. The most obvious is to establish a reliable identification key in each of the two separate systems that maintains a specified level of accuracy. Better yet, implement an individual student record data system that provides a permanent link connecting all of the student records from various data collections. Most importantly, standardized coding and verification procedures are needed within each data collection component so that performance measurement can incorporate and verify information from multiple sources.

The primary source of measurement error that we encountered was student demographics including program participation fields. Several data fields from the testing file were employed in the analyses including birth month, birth year, gender, school district code, special education program code, and standard, modified, or exempted administration. Since our ultimate goal was to examine special education student performance on the statewide reading and mathematics assessments, these fields were needed for reliable matching and for establishing subgroups for analysis. Inaccuracies in these fields left a large portion of the matched records too suspect to include in further analysis. The question of participation remained substantively unanswered due to poor data quality and lack of a unique student identifier.

Administrative policies and procedures need to play an important role when the expectation is that all students, including special education students, will participate in testing. The monitoring system recommended by Ysseldyke would allow states to evaluate whether the policies are implemented uniformly. This will be essential to answering questions of equity. The decision about when to administer the assessment under standard, modified, or exempt conditions is influenced by the reporting system and the aggregation level of the reporting. By looking at student performance within subpopulations such as Special Education or Title 1 programs, the state extends the reporting and accountability system beyond the school and district level to the program level. This simply cannot be done without improving the accuracy of demographic data collected along with the assessment. One acceptable recommendation involves establishing a specified level of data integrity required prior to reporting on student achievement within special populations. The specified level must include degree of participation and integrity of both demographic and testing responses.

Many states are facing similar challenges in assessing and reporting on all students. Some of the factors that should be considered are listed below. These are an outgrowth of recommendations that Oregon will consider in the future when expanding the capability of the assessment program to adequately assess and report on all of its students.

- Prior to reporting on the performance of all students, states may need to examine the various data collections involved. Look for a unifying system and effective means of linking data from multiple sources. Without these, it may be difficult or impossible to link data.

- State statutes may not clearly encompass all students in data collections. Conduct an analysis of how statutes affect which data will be collected, who will collect and maintain them, when they will be collected, how they will be maintained, and who reports them. Such a review might suggest some revisions to existing statutes.

- Age-to-grade translations need to be re-examined as a means of identifying special education students in the absence of grade level designations. For assessments conducted at particular grade levels (like Oregon's assessments at 3rd, 5th, 8th, and 10th grades) age-to-grade may leave out students from non-graded programs, those who started school late, or took the same grade level over a second time.

- Testing answer sheets include demographic information needed to analyze assessment results and report findings. Student bubbling may be the source of error for birth dates and program participation. A solution should be sought such as having testing proctors review answer sheets prior to returning them.

- Ensure the accuracy of fields used in the merge key. Two approaches that might be considered are electronically pre-coding answer sheets prior to testing or establishing a unique student identifier statewide, if one does not already exist.

- Investigate the circumstances surrounding the marking of exclusion codes. Descriptions of who is exempt may lack clarity or educators in the field may be marking these fields incorrectly. A follow-up review or audit of answer sheets coded modified and exempt might inform the process.

- Improve the ability to both aggregate and disaggregate summaries. Feedback on student performance is essential to improving scores. Eliminating scores from reporting may skew results and will fail to account for the progress or lack of progress among various subgroups, including special education students.

References

Algozzine, R. (1993). Including students with disabilities in systemic efforts to measure outcomes: Why ask why? In J. Ysseldyke & M. Thurlow (Eds.), *Views on inclusion and testing accommodations for students with disabilities* (Synthesis Report 7, pp. 5-18). Minneapolis, MN: National Center on Educational Outcomes.

Elliott, J., Thurlow, M., & Ysseldyke, J. (1996). *Assessment guidelines that maximize the participation of students with disabilities in large-scale assessments: Characteristics and considerations* (Synthesis Report 25). Minneapolis, MN: National Center on Educational Outcomes.

Erickson, R., Thurlow, M., & Ysseldyke, J. (1996). *Neglected numerators, drifting denominators, and fractured fractions: Determining participation rates for students with disabilities in statewide assessment programs* (Synthesis Report 23). Minneapolis, MN: National Center on Educational Outcomes.

Jaeger, R. M. (1992). *"World class" standards, choice, and privatization: Weak measurement serving presumptive policy.* Vice-Presidential Address to Division D presented at the annual meeting of the American Educational Research Association, San Franciso, CA.

Merwin, J. (1993). Inclusion and accommodation: "You can tell what is important to a society by the things it chooses to measure." In J. Ysseldyke & M. Thurlow (Eds.), *Views on inclusion and testing accommodations for students with disabilities* (Synthesis Report 7, pp. 30-34). Minneapolis, MN: National Center on Educational Outcomes.

Oregon Department of Education. (1996a). *Average daily membership attending—year ending June 30, 1996.* Salem, OR: Author.

Oregon Department of Education. (1996b). *Report of children and youth with disabilities receiving special education—Revised April 12, 1996.* Salem. OR: Author.

Reschly, D. (1993). Consequences and incentives: Implications for inclusion/exclusion decisions regarding students with disabilities in state and national assessment programs. In J. Ysseldyke & M. Thurlow (Eds.), *Views on inclusion and testing accommodations for students with disabilities* (Synthesis Report 7, pp. 35-46). Minneapolis, MN: National Center on Educational Outcomes.

SPSS Base 7.5 for Windows [Computer Software]. (1997). Chicago, IL: SPSS Inc.

Thurlow, M.L. (1997). *Highlights of accountability systems in two states that include all students with disabilities*. Paper presented at the symposium "The Challenge of Including All Students in State Accountability Systems: Alternatives for Students Excluded from Regular Assessments" (Session 1.40) at the annual conference of the American Educational Research Association, Chicago, IL.

Thurlow, M. L., Scott, D. L., & Ysseldyke, J. E. (1995). *A compilation of states' guidelines for including students with disabilities in assessments* (Synthesis Report 17). Minneapolis, MN: National Center on Educational Outcomes.

Thurlow, M.L., & Ysseldyke, J.E. (1993). *Can "all" ever really mean "all" in defining and assessing student outcomes* (Synthesis Report 5). Minneapolis, MN: National Center on Educational Outcomes.

Wineberg, H. (1997). *Population estimates for Oregon: July 1, 1996.* Portland, OR: Portland State University, Center for Population Research and Census.

Ysseldyke, J.E., & Thurlow, M. L. (1993). *Views on inclusion and testing accommodations for students with disabilities* (Synthesis Report 7). Minneapolis, MN: National Center on Educational Outcomes.

Ysseldyke, J.E., Thurlow, M. L., McGrew, K., & Shriner, J. G. (1994). *Recommendations for making decisions about the participation of students with disabilities in statewide assessment programs* (Synthesis Report 15). Minneapolis, MN: National Center on Educational Outcomes.

Ysseldyke, J.E., Thurlow, M. L., McGrew, K., & Vanderwood, M. (1994). *Making decisions about the inclusion of students with disabilities in large-scale assessments* (Synthesis Report 13). Minneapolis, MN: National Center on Educational Outcomes.

NATIONAL
CENTER ON
EDUCATIONAL
OUTCOMES

University of Minnesota • 350 Elliot Hall
75 East River Road • Minneapolis, MN 55455
612.626.1530 • Fax 612.624.0879
http://www.coled.umn.edu/NCEO

The College of Education
& Human Development

UNIVERSITY OF MINNESOTA