Inclusion of Limited-English-Proficient Students in Rhode Island's Grade 4 Mathematics Performance Assessment

CSE Technical Report 486

Lorrie Shepard, Grace Taylor, and Damian Betebenner
CRESST / University of Colorado at Boulder

September 1998

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

and

Center for Research on Education, Diversity and Excellence
University of California, Santa Cruz
1156 High Street
Santa Cruz, CA 95064
(408) 459-3500

Project 2.4 Assessment of Language Minority Students Lorrie Shepard, Project Director, CRESST/ University of Colorado at Boulder

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Numbers R305B60002 and R306A60001, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U. S. Department of Education.

INCLUSION OF LIMITED-ENGLISH PROFICIENT STUDENTS IN RHODE ISLAND'S GRADE 4 MATHEMATICS PERFORMANCE ASSESSMENT[1]

Lorrie A. Shepard, Grace A. Taylor, and Damian Betebenner
CRESST/University of Colorado at Boulder

Importance of Inclusion

State and national assessments have a prominent role in the current context of educational reform and school accountability. Assessments are needed to monitor the effectiveness of reform efforts and, in some cases, are intended as powerful levers to induce school improvement. As assessment results have taken on greater importance, so to has the question of who participates. In the past, English-language learners were often *excluded* from large-scale assessments, as were many students with disabilities, because limited English proficiency or the nature of students' disabilities prevented them from understanding questions or responding to the assessment as normally administered. Such exclusions, however, distort the accuracy of assessment results. Because average scores and the percent of students achieving proficiency standards are calculated on the basis of participating students, a state or district that excuses 10% or 12% of its students from testing reports a misleading picture of academic achievement. Differential exclusion rates can also lead to invalid comparisons among states and among school districts. More importantly, from the perspective of advocates for English-language learners and disabled students, exclusion removes these children from the accountability system and denies their rights to be full beneficiaries of educational reform efforts.

Inclusion of English-language learners in statewide assessments, then, is important both symbolically and technically. As a policy, full inclusion signals the commitment of the educational system to support the academic progress of all its students; and it ensures the representativeness of the data reported. At the same time, inclusion also creates a host of ethical, logistical, and technical

[1] The authors wish to express their gratitude to our colleagues in the Rhode Island Department of Education, Dr. Pasquale DeVito, Maria Lindia, and Dr. James Karon. We also thank Eric Barela, Timothy Weston, and Kerry Wheeler for their help with portions of the data analyses.

problems. Although there are numerous stories and documented cases of principals and teachers who raised test scores by telling low-scoring students to stay home, teachers are more often motivated to exclude students from formal assessments to protect them from the frustration and embarrassment of attempting work they can't understand. This need to protect students from the harmful consequences of assessment is felt most dramatically in systems where state tests are used to make high-stakes decisions about individual students as well as to report on statewide achievement levels. Participation of students who are not yet fully proficient in English requires some form of accommodation so that students can demonstrate their skills and knowledge of the subject without being hindered by the language of test questions or response format. Accommodations, especially translation of assessments into students' first languages, may be difficult and costly. Equally problematic is determining the appropriate accommodation for each student, depending not only on level of English proficiency but on formal schooling, proficiency in the students' first language, and the current language of instruction. Finally, there is the question of the validity of accommodations. Although accommodations are intended only to level the playing field by removing sources of difficulty irrelevant to the skills and knowledge being assessed, what little research exists suggests that assessment accommodations may sometimes alter the equivalence of the assessments and give an unfair advantage to students receiving accommodations. (Koretz, 1997; Willingham et al., 1988)

Definitions

Limited-English-proficient (LEP) is the official term used by the U.S. federal government to designate students whose first language is not English and who lack the English skills to receive instruction only in English. LEP is the term used in Rhode Island assessment materials when teachers are asked to consider inclusion rules and possible accommodations for English-language learners. In this report, we use the term LEP when referring to procedures and data from the Rhode Island assessment. In our more general discussions, however, we use the term *English-language learner*, first proposed by LaCelle-Peterson and Rivera (1994) to focus "on what students are accomplishing, rather than on any temporary 'limitation' they face" (p. 55). This usage is modeled after the terminology in a recent report of the National Research Council (NRC), *Improving Schooling for Language-Minority Children* (August & Hakuta, 1997).

Note that *language-minority* student is a more general term that includes both English-language learners and *non-native speakers of English* who are now proficient in English and hence *bilingual*. A more complete taxonomy is provided by Butler and Stevens (1997).

Accommodations are adaptations or changes in how an assessment is administered or in the mode of response. The intention of accommodations is to remove irrelevant sources of difficulty, to get a fairer or more accurate picture of what the test-taker actually knows. For example, if an assessment is intended to measure students' knowledge and problem-solving abilities in mathematics, then testing English-language learners in English may not allow them to demonstrate the full extent of their mathematical understandings. In another recent National Research Council report addressing instructional and assessment issues affecting students with disabilities (McDonnell, McLaughlin, & Morison, 1997), assessment accommodations were likened to the use of a corrective lens. "Testing accommodations are intended to offset or 'correct' for distortions in scores caused by a disability" (p. 249). In the case of students with disabilities, testing accommodations may included Braille and large-print versions of the test for students with visual disabilities, scribes for students who are not physically capable of writing, and small-group settings or extra time for students with learning disabilities. For English-language learners, accommodations include test translation, oral reading of the test in English, and use of dictionaries, as well as extended testing time. Figure 1, from Butler and Stevens (1997), shows the types of assessment accommodations made for English-language learners. Figure 2 is taken directly from the 1997 Rhode Island Assessment Program materials and shows in detail what accommodations were available either for LEP students or for students with disabilities. The only type of accommodation suggested by Butler and Stevens that was not available in Rhode Island was an adaptation of the vocabulary or linguistic complexity of the test.

Assessment Research Framework

A central principle of validity theory is that validity depends on test use. This means that any investigation of a test's validity, for students generally or for English-language learners, must be undertaken in the context of specific assessment applications. The assessment research framework presented in Figure 3 creates a structure for identifying the main content domains and

Two Categories of Accommodations for English Language Learners	
Modifications of the test	Modifications of the test procedure
• Assess in the native language • Text changes in vocabulary • Modification of linguistic complexity • Addition of visual supports • Use of glossaries in native language • Use of glossaries in English • Linguistic modification of test directions • Additional example items	• Extra assessment time • Breaks during testing • Administration in several sessions • Oral directions in the native language • Small-group administration • Separate room administration • Use of dictionaries • Reading aloud of questions in English • Answers written directly in test booklet • Directions read aloud or explained

Figure 1. Potential accommodation strategies for English-language learners (Butler & Stevens, 1997).

categories of assessment purpose, which must be considered when designing research studies. This structure serves as a road map to locate research on accommodations within a larger set of topics dealing with assessment of English-language learners and bilingual students. The 3 x 4 matrix presents three different *assessment purposes*:

⇒ use of assessment for instructional planning within the classroom,

⇒ system-level monitoring and accountability, and

⇒ program placement or exit,

and four *assessment domains*:

⇒ subject-matter knowledge,

⇒ native language and literacy,

⇒ English language and literacy,

⇒ cognitive abilities.

Codes starred once (*) are NOT applicable in the Writing Assessment.
Codes starred twice (**) are ONLY applicable in the Health Assessment.
(Enter appropriate code(s) in the Testing Accommodations section of Student Information Sheet.)

Administration Accommodations

Code #

01 Braille edition of assessment
02 Large-print edition of assessment
03 Use of magnifying equipment
04 Oral reading of assessment
05 Signing of assessment
06 Repeated directions
07 With student using amplification equipment (e.g., hearing aid or auditory trainer)
08 Written translation of assessment into Spanish
09 Oral administration of test in Spanish
16 Oral administration of assessment in another language (*Specify on the Supplementary Form*)
17* Use of translation dictionaries
18* Using visual aids
19 Other accommodation (*Specify on the Supplementary Form*)

Response Accommodations

Code #

20 Use of typewriter for responding
21 Use of computer/word processor for responding
23* Giving response orally (written verbatim by test administrator)
24** Giving response orally to a tape recorder
25* Giving response in sign language (written verbatim by test administrator)
26* Writing response in Spanish
27* Giving response orally in Spanish (written verbatim by test administrator)
28** Giving response orally in Spanish (written verbatim by test administrator)
35* Writing response in another language (*Specify on the Supplementary Form*)
36* Giving response orally in another language (written verbatim by test administrator) (*Specify on the Supplementary Form*)
37** Giving response orally in another language to a tape recorder (*Specify on the Supplementary Form*)
38* Adult transcription of portion of student's writing
39 Other accommodation (*Specify on the Supplementary Form*)

Setting Accommodations

Code #

40 Testing in special education or resource classroom
41 Testing with small group
42 Testing individually
43 With the student seated in front of classroom
44 With teacher facing student (hearing impaired)
45 Testing in ESL classroom
49 Other accommodation (*Specify on the Supplementary Form*)

Timing Accommodations

Code #

50 Extended time (if testing exceeds 10 minutes beyond recommended time, either or both days)
51 More frequent breaks during testing
52 Extended testing sessions over several days
59 Other accommodation (*Specify on the Supplementary Form*)

Not Able to Accommodate

Mark code "Not tested" and "Reason for Not Testing" in Location L of the Student Information Sheet.
This should be marked only if none of the above accommodations would assist this student in successfully completing this assessment (Also complete the Supplementary Accommodation Information Form.)

Figure 2. Rhode Island State Assessment Program, Spring 1997, summary of health, mathematics and writing performance assessment accommodations.

	Assessment Purpose		
Proficiency Domain	Instruction	System-Level Monitoring & Accountability	Program Placement & Exit
Subject Matter Knowledge	1	2 Academic Achievement	3
Native Language & Literacy	4	5	6
English Language & Literacy	7	8	9 Eligibility for ESL Services
Cognitive Abilities	10	11	12 Special Education Identification

Figure 3. Research framework for assessment of language-minority students.

The present study is located in Cell 2 of the matrix. It addresses accommodations and validity issues in the context of large-scale assessment programs designed to assess students' *subject-matter knowledge* (e.g., mathematics) for purposes of *system-level monitoring and accountability*. Two other cells in the matrix are highlighted. Cell 9 refers to assessment of students' English language proficiency to determine eligibility for English-as-a-second-language services as well as to exit students from such programs, and Cell 12 refers to assessment of students' cognitive functioning as part of an evaluation for placement in special education programs. These other two categories are important because, until very recently, most research on assessment of English-language learners has focused on these assessment purposes, which are entirely different from assessment of students' content knowledge as part of a large-scale assessment.

Even for these two categories of assessment practice with a longer history, the NRC summary of research presents a gloomy picture of the current state of knowledge (August & Hakuta, 1997). For example, existing English-language proficiency instruments measure a limited range of language skills and are

inconsistent with more contemporary models of first- and second-language acquisition and literacy development. For the purpose of evaluating potential learning disabilities, there are no instruments available that can adequately disentangle evidence of disability from the confounding effects of second-language learning. Although there are promising dynamic assessment techniques that evaluate students' learning potential only after providing structured learning opportunities, assessment personnel are not trained in these methods and generally lack expertise in evaluating linguistically and culturally diverse learners (August & Hakuta, 1997).

The assessment research framework was devised by Shepard (1995) to guide research in the future, especially to emphasize the substantive parallels between instructional and accountability assessments. Although classroom-level and state-level assessments invoke very different practical and technical issues, and therefore require distinct research studies, there should nevertheless be a close *substantive* linkage between the content of these two types of measures. Shepard (1996) suggested that research should focus on "conceptualizing and developing performance continua in each proficiency domain," to which both teacher-based and system-level assessments could be anchored. Such conceptual mappings of students' developing proficiencies, illustrated with benchmark samples of student work, would support the learning of all students but would also provide a basis for modeling the increasing subject-matter knowledge of English-language learners as they more and more closely approximate common performance standards.

In addition to research aimed at documenting academic proficiencies as they develop over time, it will also be critical to study how such patterns are mediated by students' particular settings and experiences. Butler and Stevens (1997) have developed a model that identifies the sociocultural and personal factors affecting the academic achievement of English-language learners, which therefore must be considered when assessing achievement, whether at the classroom or system level. Elements in their model include community factors, such as ethnic diversity, language use, community attitudes toward immigration and language differences, and the socioeconomic status of the neighborhood; school factors, such as the quality and types of programs, student opportunity to learn, teacher training and background, and classroom discourse practices; and home factors, which include parent educational background, home literacy practices, and

parental beliefs and involvement with their child's education. Individual student factors affecting learning include personal characteristics, such as motivation, attitudes toward American culture, age of arrival in the U.S., and length of time since arrival; educational background, especially years of formal schooling and quality of instruction in the student's home country; and language factors, such as native language proficiency, academic language proficiency in English, and the language of instruction.

These conceptual models lay out an ambitious research agenda; yet against this backdrop the present study can be regarded as only exploratory. Research on the use and validity of accommodations for English-language learners is just beginning, and as is evident in the next section, is still at a very crude and simplistic stage compared to the complexity of issues.

Previous Research on Accommodations in Large-Scale Assessments

A 1997 report by the National Center for Education Statistics (Olson & Goldstein) offers a useful summary of research to date on *The Inclusion of Students With Disabilities and Limited English Proficient Students in Large-Scale Assessments*. The report also provides an overview of technical issues and studies currently underway. Given the recency of efforts to increase the participation of English-language learners, it is not surprising that most studies are descriptive rather than evaluative. For example, it is a nontrivial task merely to estimate the number of limited-English proficient students in the U.S. (approximately 2.3 million, 5.5% of the U.S. student population; Fleischman & Hopstock, 1993) and to document the distribution of such students by native language and by state. Fleischman and Hopstock (1993) found that 72.9% of LEP students speak Spanish as their primary language. The next most frequent language is Vietnamese, spoken by 3.9% of LEP students. According to the 1990 U.S. Census, 30% of children in California ages 5-17 were reported to speak a language other than English in the home and were rated as speaking English less than "very well." Sixty-seven percent of language-minority students live in five states—California, Texas, New York, Florida, and Illinois.

Other descriptive studies report the extent of inclusion and exclusion practices as well as the use of various types of accommodation by state assessment programs. For example, an important finding of the Council of Chief State School Officers and North Central Regional Educational Laboratory (1996) survey was that most states permitted exclusion of LEP students, usually based

on a language proficiency measure or number of years in the U.S. Also most states provide accommodations for LEP students who do participate in assessments, but these accommodations more frequently involve a change in the assessment administration—separate testing session, flexible scheduling, small-group administration or extra time—rather than a change specifically focused on the language demands of the assessment. Nine states allowed the use of dictionaries or word lists as an accommodation. Only five states translated tests or developed tests in languages other than English. The tendency to focus accommodations on test setting and time limits makes sense if the only language-minority students participating are those with some degree of English proficiency. As states move to full inclusion, however, assessing students with little or no English proficiency would require translation or other changes in the linguistic demands of both the assessment and mode of response.

A few in-depth studies have been undertaken to examine how exclusion practices might affect assessment results. Stancavage, Allen, and Godlewski (1996) conducted individual Spanish-language assessments of LEP students sampled as part of the 1994 NAEP Trial State Assessment in reading. Despite the NAEP directions to be as inclusive as possible, a surprising finding was that more than three quarters of the excluded students had spent four or more years in English-speaking settings. Furthermore, when Spanish-bilingual site visitors proceeded *in English* to administer individualized reading assessments using a second-grade story followed by a block of 4th-grade NAEP reading items, the researchers judged that more than 75% of the excluded LEP students could have participated in the assessment. In reaching this conclusion, the National Academy of Education panel overseeing the study acknowledged that language factors undoubtedly caused the assessment to underestimate the true reading proficiency of some LEP students but argued that "estimates of student achievement need only be accurate enough to allow scores for these students to contribute to state averages, not to make conclusive judgments about the achievement of individual students" (National Academy of Education, 1996, p. 67). It is likely that teachers and researchers were operating from very different perspectives in this regard. Teachers were much more liberal than researchers in recommending both accommodations and exclusions, probably because they were reasonably striving to prevent student achievement from being underestimated.

Experimental studies of the kind discussed in the next section, designed to evaluate the effects of accommodations on performance, are almost non-existent. Abedi, Lord, and Plummer (1997) observed that language-minority students performed more poorly on NAEP mathematics items that required an extended response or that involved complex language structures or unfamiliar vocabulary. They followed up with a randomized experiment, comparing performance of English-language learners on originally worded items versus equivalent items with simplified wording. The study showed that reducing language complexity of items improved the performance of English-language learners in low- and middle-level math groups.

There have been more experimental studies evaluating the effects of accommodations on performance for students with disabilities than for English-language learners. However, these studies have been in the context of college entrance examinations, particularly the SAT. Although it would be a mistake to generalize findings from students with disabilities to English-language learners, findings from college admissions accommodations do sound a cautionary note. Contrary to the intention of increasing validity of test results by removing only irrelevant sources of difficulty, in controlled studies accommodations provided on the SAT and GRE reduced rather than increased the predictive validity of test results (Willingham et al., 1988). In particular, providing extra time appeared to give too much of an advantage to students with disabilities and led to overprediction of college GPAs (Braun, Ragosta, & Kaplan, 1988). Findings like these make it clear that the effects of accommodations on assessment validity cannot be taken for granted, and they point to the kinds of comparative studies needed to evaluate both performance effects and validity.

One additional study deserves mention because it addressed accommodations for students with disabilities in a large-scale assessment program and because its exploratory nature was very much like the present study. Koretz (1997) investigated accommodations for students with disabilities in the Kentucky state assessment in Grades 4, 8, and 11. An important feature of the Kentucky context was the extensive effort made to be as inclusive as possible. In fact, more than 80% of students with disabilities were assessed, and most of these were provided with two or more accommodations. Koretz termed his findings mixed regarding the psychometric effects of accommodations. Analogous to findings for college admissions tests (Willingham et al., 1988),

internal correlational and structural analyses indicated that the assessments seemed to be measuring in similar ways for students with and without disabilities. Other findings, however, raised questions about the credibility and validity of results. Koretz cited the high frequency of accommodations, especially in the fourth grade, as a sign of possible misuse. More seriously, the high scores of learning-disabled students and mentally-retarded students receiving certain types of accommodation seemed implausible given that students in these groups would not be expected to be above average in performance.

Needed Research on Accommodations

Three important lessons can be learned from the existing research on accommodations. First, the corrective lens provided by accommodations may not work as intended. Second, improved performance might not be evidence of improved validity. A third lesson, a methodological one, should also be apparent. It is difficult to evaluate the effects of accommodations in the context of operational assessment programs because it is not possible to compare how any given student would have done without the accommodation. The results that Koretz observed, for example, were interpretable only because they were so far out of line. If mentally retarded students had turned in below-average performances, researchers would not have known if results were valid or inflated. Controlled studies are needed to evaluate whether accommodations correct an unfair disadvantage or overcompensate in a way that reduces the validity of assessment results. The ideal study for most accommodations is a 2 x 2 experimental design with both English-language learners and native speakers of English being randomly assigned to both accommodated and non-accommodated conditions. This design would work, for example, to study the effects of extra time or of providing dictionaries (two-way dictionaries for English-language learners and English dictionaries for monolingual speakers). Other study methods would be needed to evaluate the equivalence of translated assessments.

If assessment accommodations are working as intended, the results should show an interaction effect. The accommodation should improve the performance of English-language learners but should leave the performance of native-English speakers unchanged. If accommodations such as extra time, small-group sessions, or repeating directions improve the performance of both groups, then providing the accommodation only to English-language learners is potentially unfair. Before deciding whether to alter assessment conditions for all

students, however, validity data should be evaluated. In controlled studies, more in-depth data should be collected through individualized assessments or classroom observations to serve as criterion measures of student achievement. Accommodations should increase the correspondence between assessment results and validity criteria for English-language learners. These data are key to answering whether improved performance has increased validity. Given that validity correlations are generally lower for English-language learners than for other groups (even after accounting for the restricted range of performance), an accommodation that would benefit everyone could be given only to English-language learners if it differentially improved validity for this group.

As we demonstrate in this study of the Rhode Island assessment program, it is possible to gather validity evidence concurrent with an operational assessment program. Concurrent validity data can be used to evaluate whether an assessment appears to be as valid for English-language learners as it is for native-English speakers. However, just as when trying to determine the effect of accommodations on average performance, controlled studies are needed to determine whether accommodations improved validity compared to the same assessment without accommodations.

Beyond simple comparative studies, Butler and Stevens (1997) have outlined a research agenda aimed at improving the match of specific accommodation to student needs and thereby building in greater validity. Their model, described earlier, would be used to identify sociocultural and personal factors that account for differences in the effectiveness of accommodations (again where improved validity would be evaluated by experimental comparisons but with groups assigned to the most appropriate accommodation). For example, providing dictionaries is likely to be more effective for students who have higher levels of English proficiency. Written translations of assessments are likely to be effective for students who received formal schooling in their native language, whereas students without formal schooling might benefit most from an oral administration of a translated version. Ultimately findings from these kinds of studies would have to be turned into simple decision rules that would match English-language learners to the appropriate accommodation. Stevens, Butler, and others are working to devise a measure of academic language proficiency that would aid in this process. Even with such improvements, however, it is unlikely that standardized decision rules will be able to capture the full

complexity of how language learning and academic learning interact, and as a result, the achievement of English-language learners will continue to be misrepresented by external assessments. Therefore, other non-experimental studies will also be needed to examine other possibilities such as using benchmarking of classroom-level assessments to link with external accountability assessments.

The Rhode Island Grade 4 Mathematics Performance Assessment Study

The newly developed Rhode Island Performance Assessment program is a particularly fruitful site for investigating accommodations for several reasons. First, the additional language demands of performance assessments make the issue of accommodations even more important than in traditional testing programs. Second, Rhode Island, like Kentucky, is further along than many states in establishing a policy of full inclusion for its statewide assessment. Third, the Rhode Island Department of Education administers the Metropolitan Achievement Test (MAT) in addition to the performance assessment, which makes it possible to compare relative performance on two very different types of measures.

Fourth-grade mathematics. The Rhode Island State Assessment Program includes performance assessments in writing, health, and mathematics. Students are assessed in Grades 4, 8, and 10. In mathematics at Grades 8 and 10, the New Standards Reference Examination in Mathematics is administered rather than a performance assessment developed by Rhode Island. Fourth-grade mathematics was selected as the subject area and grade level for this study because mathematics is the content area where students can most clearly develop content knowledge independent of their English language proficiency and because the proportion of English-language learners is greatest in the elementary grades.

Assessment instruments. In spring 1997, the Rhode Island Grade 4 performance assessment in mathematics was administered in two 60-minute sessions on two separate days, with an additional 10 minutes allowed if students in a class were still working on the assessment at the end of the hour. Each student completed 10 multi-part problems scored using a 0-4 rubric. Because two problems were matrix sampled and varied from student to student, only the 0-32 scores based on common problems were used for analysis. Problems included matching a story to data in a graph, estimation, multiplication and division applications, representing numbers with base ten stickers, and representing

tangrams with numbers. All of the problems required students to explain their answers.

The Metropolitan Achievement Test (Elementary 2, Form S; Balow, Farr, & Hogan, 1993) has two mathematics subtests, Concepts and Problem Solving, and Procedures. Because administration of the Procedures section was not required, and therefore had very low participation rates, only the Concepts and Problem Solving subtest was used for analysis. Items required students to read information from tables, interpret fractions, complete number sentences, identify information needed to answer questions, and solve word problems.

Inclusion and accommodations. Directions to teachers and school administrators for administering the Rhode Island State Assessment Program emphasize that "all students are expected to participate in the performance assessments." Special consideration for LEP students were identified as follows:

- Performance Assessments should be given in the language in which the student is most capable of showing knowledge and skills. Either a written translation or an oral administration (in the native language or in English) may be used.

- In considering language accommodations, think about each of your LEP student's
  - amount of formal schooling in their country of origin,
  - amount of schooling in the U.S., and age when he/she came to the U.S.

- For mathematics, a scribe will be needed for oral responders, or students may write their responses in their native language.

The Rhode Island Department of Education provided special testing materials in Spanish and assisted districts in identifying bilingual interpreters and scribes. As guidelines, it was suggested that fewer than 2% of all students enrolled would be expected to be unable to participate in the assessments, and it was expected that 7% to 10% of students would require one or more accommodations. These figures referred to LEP and students with IEPs combined. The range of accommodations provided is shown in Figure 2.

Special pilot study. In addition to statewide results for fourth graders on both the Mathematics Performance Assessment and Metropolitan Achievement Test, classroom-level data were collected from a sample of 22 volunteer classrooms with significant numbers of English-language learners. Teachers were

asked to provide additional information about students to be used in evaluating the validity of the assessments. Teachers first listed all of the students in their classes in quartile groupings ("Students in the top quarter of the class," "Students in the next-to-top quarter of the class," and so forth) and then recorded first-semester and third-quarter mathematics grades for each student. They also gave a standards-based rating of mathematics achievement (Below Basic, Basic, Proficient, or Exemplary) using the Rhode Island definitions of each proficiency level, and a language-proficiency rating using the scale shown in Figure 4. Using the quartile groupings, teachers identified target students in each group and collected examples of mathematics assignments completed by these students during May of 1997. Teachers were asked to identify one native speaker of English from each quartile and up to three non-native speakers of English from each quartile.

Participation Numbers and Percents

Data in Table 1 report the numbers of fourth-grade students who participated in the Mathematics Performance Assessment and in the Metropolitan Achievement Test. Two different sets of numbers are given for the performance assessment. The first is the total number of students accounted for in the data set, including 400 students who did not take the assessment but for whom teachers completed data records. The second indicates the number of students who actually took the assessment. To provide a basis of comparison for

Write "Mono" for Monolingual Speakers of English. This student is a native speaker of English.

Write "5" for Level 5 Advanced Student. This student is *not* a native speaker of English but is verbally proficient in English. This student no longer receives E.S.L. services but may still be monitored.

Write "4" for Level 4 Advanced Intermediate. This student is continuing to gain fluency in English but is in the refinement stage.

Write "3" for Level 3 Intermediate. This student is working on increasing verbal ability and is at the expansion stage.

Write "2" for Level 2 Advanced Beginner. This student is transitioning from a silent period and is at the developmental stage for expressive/receptive language.

Write "1" for Level 1 Beginner. This student may be in a silent period and has no or minimal receptive/expressive language skills in English.

Figure 4. Language-proficiency rating scale.

Table 1

Numbers of Grade 4 Students Who Took the Mathematics Performance Assessment (PA) and Metropolitan Achievement Test (MAT) and Percents of Statewide Enrollment

	General education	LEP < 2 years	LEP ≥ 2 years	Special education ≥ 50%	Special education < 50%	All students
Total numbers of students enrolled	11,129	882		2,319		14,330
Total number of students in PA data set	9,903 89%	162 18%	572 65%	514 22%	968 42%	12,042 84%
Students with PA scores	9,673 87%	139 16%	554 63%	412 18%	938 40%	11,642 81%
Students with MAT scores	9,740 88%	48 5%	552 63%	206 9%	882 38%	11,378 79%
Students with both MAT and PA scores	8,348 75%	43 5%	421 48%	173 7%	782 34%	9,926 69%

evaluating participation rates, state enrollment data (October 1996) are also reported for general education students, LEP students, and students with an individualized education plan (IEP).

One striking finding is that statewide only 84% of all fourth-graders were accounted for in the Mathematics Performance Assessment data. Although some of this nonparticipation is due to students' disabilities or language proficiency, this effect must also be due to absences on the days of testing and perhaps the more generous way that student enrollments are counted for census purposes, because even among general education students who were neither LEP nor in special education, only 89% were accounted for in the performance assessment data set.

Nonparticipation was greater for limited-English proficient students and for students in special education. Eighty-three percent of LEP students statewide were accounted for in the performance assessment data set; 79% actually took the assessment. This was a significant increase in participation by LEP students compared to the Metropolitan where only 68% took the test. Not surprisingly, all of this increase in participation occurred for LEP students who had less than 2

years of education in the U.S. This difference was very likely attributable to the availability of accommodations on the performance assessment, which were not available on the Metropolitan. The pattern for special education students was similar, but overall there was much lower participation for students with disabilities than for LEP students. Statewide, only 64% of the total special education enrollment was accounted for in the performance assessment. Only 58% actually took the assessment but this was an improvement over the Metropolitan where only 47% of special education students participated. Again, almost all of the gain in participation was with the more seriously affected group.

Additional data are provided for the matched data set of students who had scores on both the Mathematics Performance Assessment and the Metropolitan. These data are used in subsequent analyses so it is important to note how the matching constraint may have altered the representativeness of the data. The greatest loss of data occurred in the general education category and in the LEP group with two or more years of education in the U.S. General education students dropped from 87% of students having scores to only 75% of enrolled students having scores on both tests. LEP students dropped from 68% of enrolled students taking the Metropolitan to 53% taking both tests. As discussed below, attrition due to matching tended to raise average scores slightly, but the effects were quite small.

Performance Levels for Students With and Without Accommodations

Statewide assessment results are reported in Table 2. Overall, Rhode Island fourth graders perform well compared to national norms. On the Metropolitan Concepts and Problem Solving subtest, the scale score mean of 605 is equivalent to the 55th percentile. This result could be slightly inflated given that Rhode Island was just at the national average on the 1996 National Assessment of Educational Progress for Grade 4 Mathematics (220 average scale score versus 222 for the nation).

The results also reveal tremendous variability among groups. General education students—those who were not identified as either LEP or in special education—scored at the 62nd percentile nationally on the Metropolitan, whereas LEP students with less than 2 years in the U.S. or with 2 or more years in the U.S. were respectively at the 7th and 12th percentiles. Students in special education placements for 50% or more of the day and those in for less than 50% of the day scored respectively at the 6th and 25th percentiles on the Metropolitan.

Table 2

Statewide Means and Standard Deviations for Fourth Graders on the Mathematics Performance Assessment and Metropolitan Achievement Test in Mathematics

	General education	LEP < 2 years	LEP ≥ 2 years	Special education ≥ 50%	Special education < 50%	All students
Performance Assessment (PA)	15.91 7.01 (*n* = 9,763)	8.32 4.97 (*n* = 139) [-1.08]*	9.47 6.10 (*n* = 554) [-.92]*	8.33 6.16 (*n* = 412) [-1.08]*	11.90 6.72 (*n* = 938) [-.57]*	14.98 7.23 (*n* = 11,642)
PA, matched data set	16.24 6.95 (*n* = 8,348)	8.53 5.51 (*n* = 40)	10.00 5.92 (*n* = 378)	8.42 5.59 (*n* = 173)	12.08 6.80 (*n* = 782)	15.41 7.13 (*n* = 9,926)
Metropolitan Achievement Test (MAT)	611.52 40.67 (*n* = 9,740)	552.67 27.91 (*n* = 48) [-1.45]*	562.88 33.12 (*n* = 552) [-1.20]*	549.08 38.28 (*n* = 206) [-1.54]*	576.58 35.98 (*n* = 882) [-.86]*	605.36 42.83 (*n* = 11,378)
MAT, matched data set	613.47 40.42 (*n* = 8,348)	558.40 32.29 (*n* = 40)	562.43 31.21 (*n* = 378)	552.55 35.35 (*n* = 173)	578.99 36.42 (*n* = 782)	606.75 42.67 (*n* = 9,926)
PA students with accommodations	14.76 7.30 (*n* = 705)	7.59 4.94 (*n* = 94)	9.84 6.11 (*n* = 383)	8.14 6.11 (*n* = 348)	12.90 6.66 (*n* = 471)	12.02 7.16 (*n* = 1,943)
PA students w/o accommodations	16.00 6.98 (*n* = 8,968)	9.87 4.73 (*n* = 45)	8.65 6.02 (*n* = 171)	9.31 6.35 (*n* = 64)	10.89 6.64 (*n* = 467)	15.57 7.10 (*n* = 9,699)

* Effect sizes were calculated by subtracting the general education mean from the subgroup mean and then dividing the difference by the standard deviation of the general education population.

Overall, accommodations appeared to improve the performance of both LEP students and students with disabilities. This can be seen by examining how far each group is below the general education mean on the performance assessment compared to the corresponding gap on the Metropolitan. LEP students with less than 2 years in the U.S. were 1.45 standard deviations below the general education mean on the Metropolitan; but despite including many more of these students on the performance assessment, they were as a group only 1.08 standard deviations below the general education mean. Effect sizes for group differences compared to the general education mean are shown in brackets. In each case, the LEP and special education groups were less far behind on the performance assessment than on the Metropolitan.

Unlike the findings reported by Koretz (1997), the improvement in relative performance attributable to accommodations does not appear to have greatly inflated scores. Reading across Table 2 the respective groups are 1.08, .92, 1.08, and .57 standard deviations below the general education mean on the performance assessment. Although it is not possible to evaluate whether these performance levels are a valid reflection of students' true proficiencies, the pattern at least seems reasonable. For example, students who are in special education less than 50% of the day would be expected to be below average in performance but not as far below as those with more serious cognitive and behavioral disabilities. LEP students do not have cognitive disabilities, but the effects of language learning concurrent with academic learning can depress performance at least initially. By definition, students identified as LEP, even those who have had 2 or more years of education in the United States, still have sufficient language needs to require special services. Students for whom English is a second language but who are deemed proficient are likely to be achieving at higher levels but are not identifiable in the data set as a group distinct from monolingual English speakers.

Variability in Use of Accommodations

The data in Table 2—intended to show the relationship between accommodations and performance levels—also raised interesting questions about the use of accommodations. In the last two rows of the table, performance assessment results are disaggregated for students with and without accommodations. While the overall pattern of results in column 1 "makes sense" in terms of the severity of language need and disability, this pattern did not hold true for students in severely affected groups who did not receive accommodations. First, it is surprising that 45 LEP students with less that 2 years of education in the U.S. received *no* accommodations. Similarly, 64 special education students who were in special education more than 50% of the day received *no* accommodations. More surprising, however, was the higher performance levels of these groups (LEP < 2 years, Special Ed. ≥ 50%) without accommodations compared to similarly classified students with accommodations. We would expect accommodations to improve relative performance, not lower it. The only possible explanation that would account for these results is if students without accommodations were a select sample, for example, they could be immigrant children with high levels of academic

preparation in their native country or students with physical handicaps that do not affect cognitive functioning. Whatever the explanation, it does not hold true for the two more mildly affected groups where, indeed, accommodations are accompanying by higher rather than lower performance levels.

To find out more about why some students, who appeared to be most in need of accommodations, did not receive them, additional analyses were conducted on the use of accommodations across schools. In Figures 5 and 6, the number of schools accommodating various proportions of LEP and special education students is shown. For example, of the 69 elementary schools with

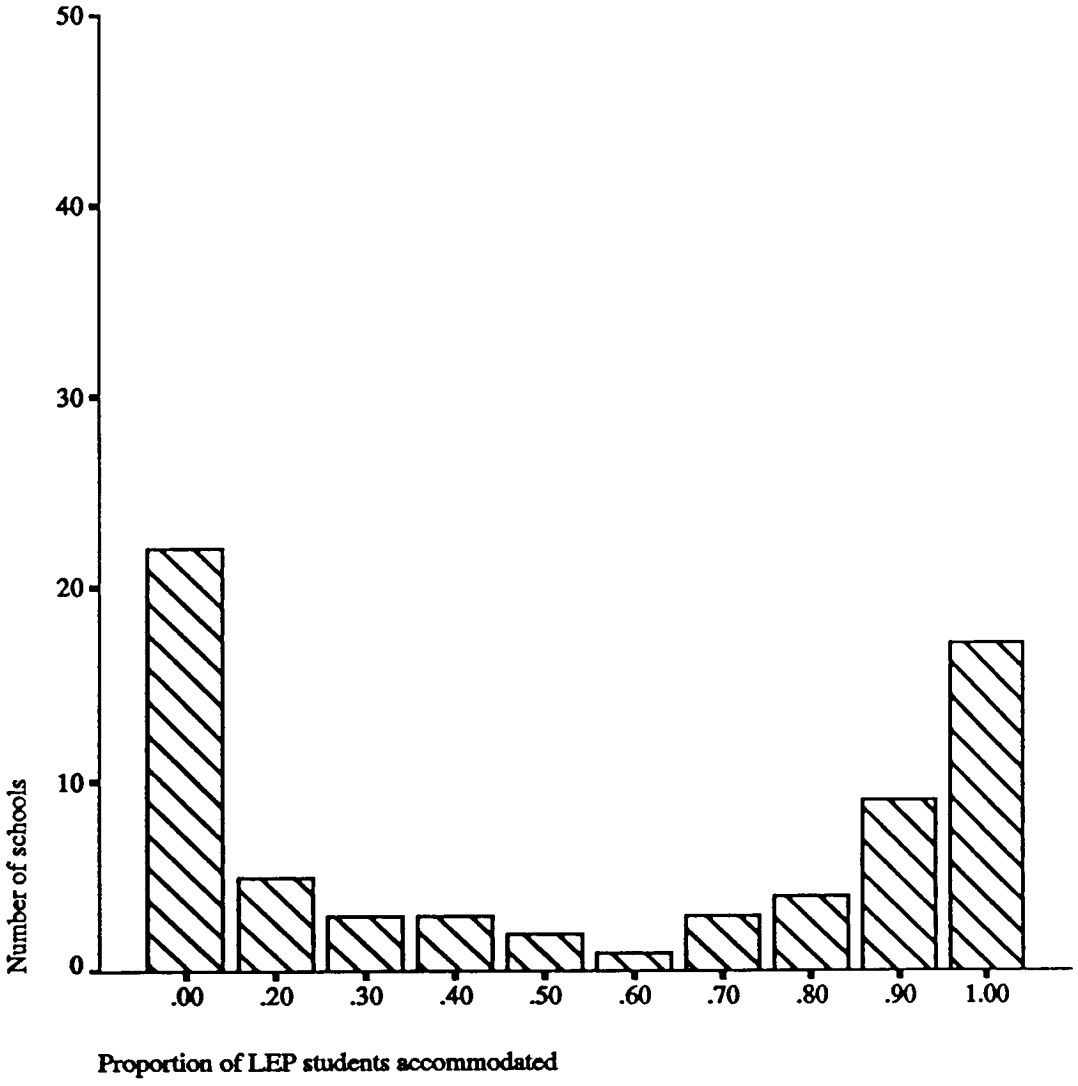

Figure 5. Number of schools accommodating various proportions of LEP students (n = 69).

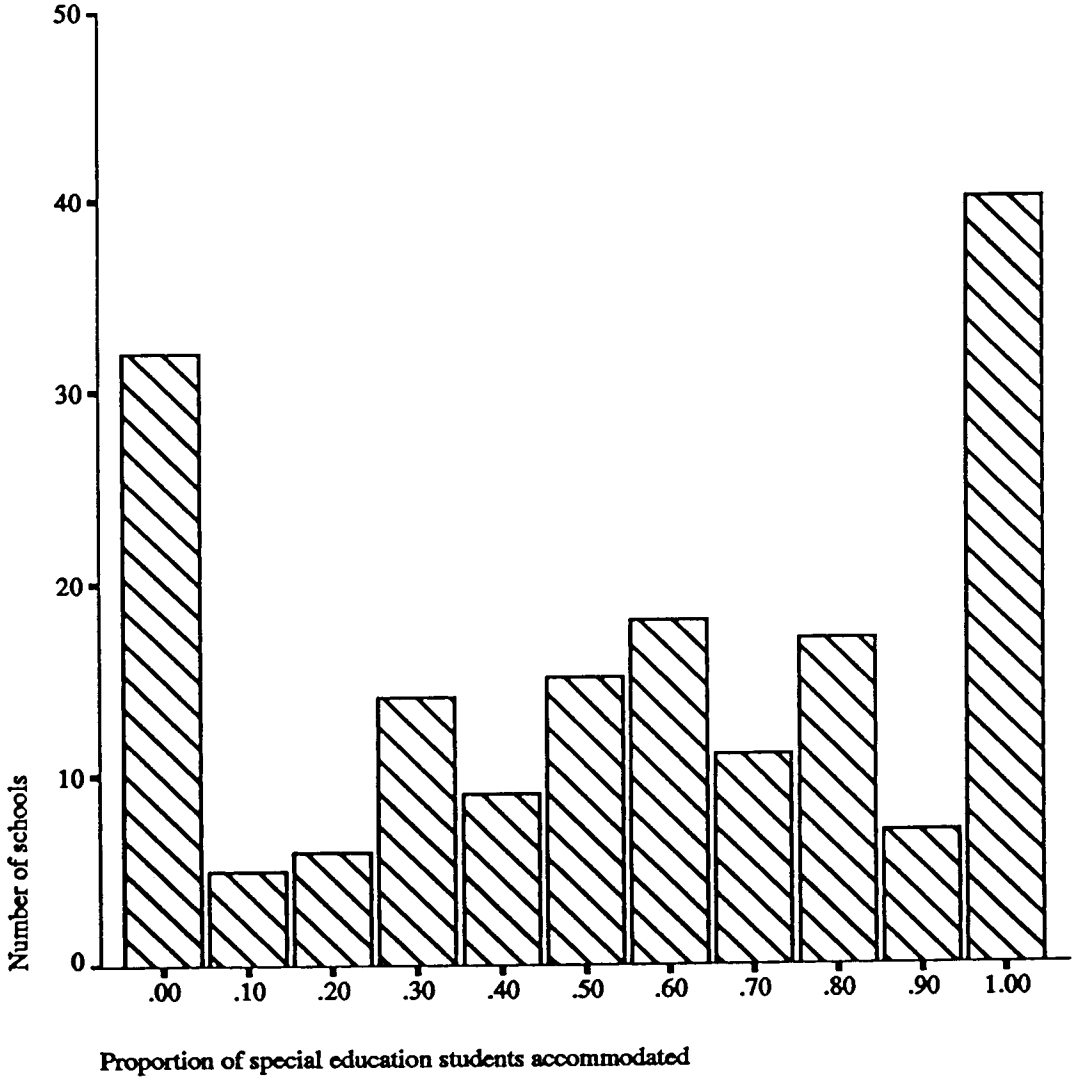

Figure 6. Number of schools accommodating various proportions of special education students (n = 174).

LEP students, 22 schools accommodated none of their LEP students, and 17 schools accommodated all of their LEP students. As illustrated by the "U" shaped distributions in both Figures 5 and 6, these two extremes were more frequent than the practice of individualizing accommodation decisions for LEP or special education students within a school. Thirty-two schools provided no accommodations to any of their special education students.

These school-level patterns in the use of accommodations were analyzed further by comparing schools with 10 or more LEP students versus those with less than 10 LEP students per grade (Figures 7 and 8), and similarly those with 10

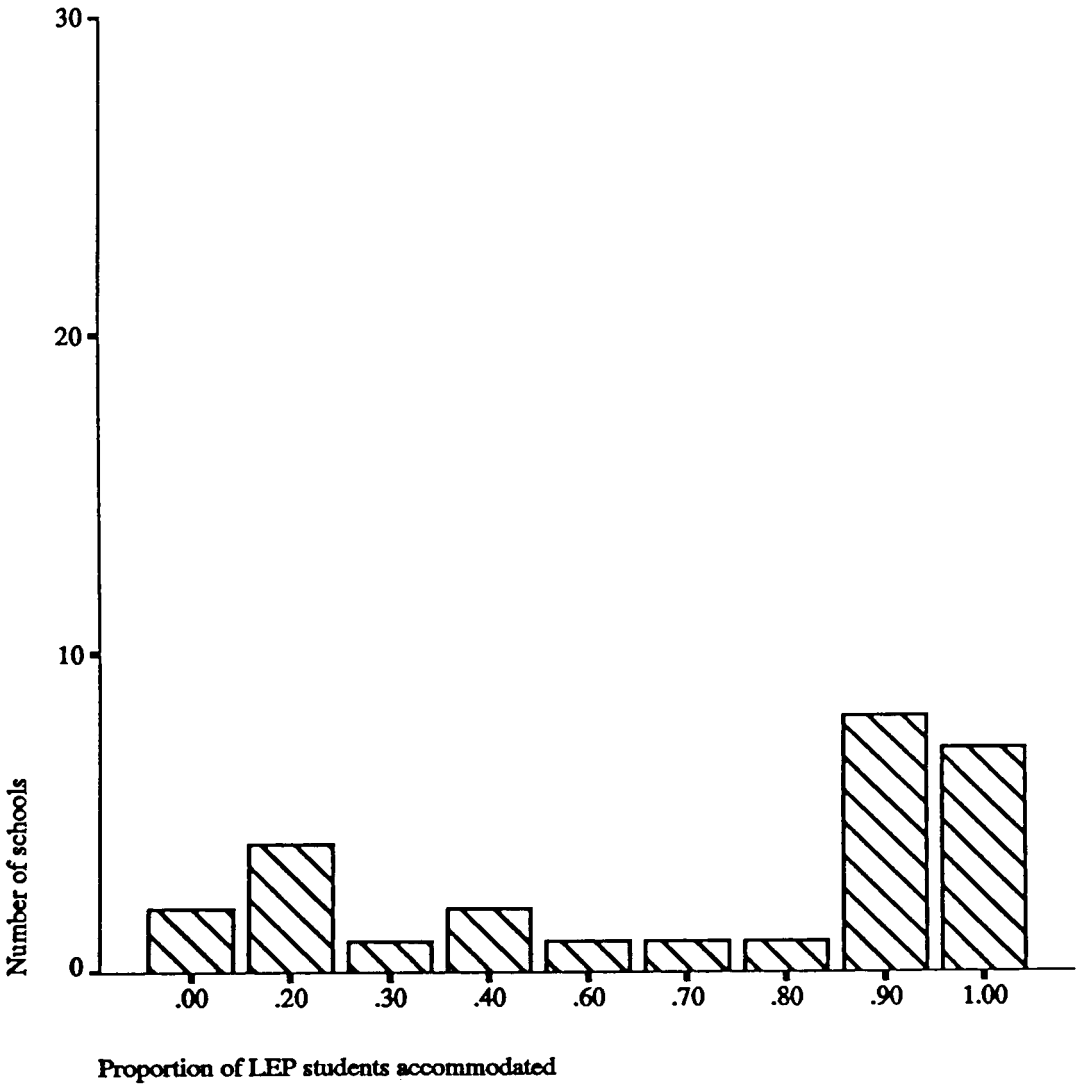

Figure 7. Number of schools, with ten or more LEP students, accommodating various proportions of LEP students ($n = 27$).

or more special education students compared to those with less than 10 such students per grade (Figures 9 and 10). Not surprisingly, schools with a larger number of LEP students reported accommodating 90% or 100% of their LEP students. Whereas, schools with fewer LEP students were less likely to provide accommodations. For special education students in schools with few such students, the pattern was again a "U-shaped" distribution. A large number of schools with few special education students provided no accommodations, but an equally large number of these schools reported accommodating all of their

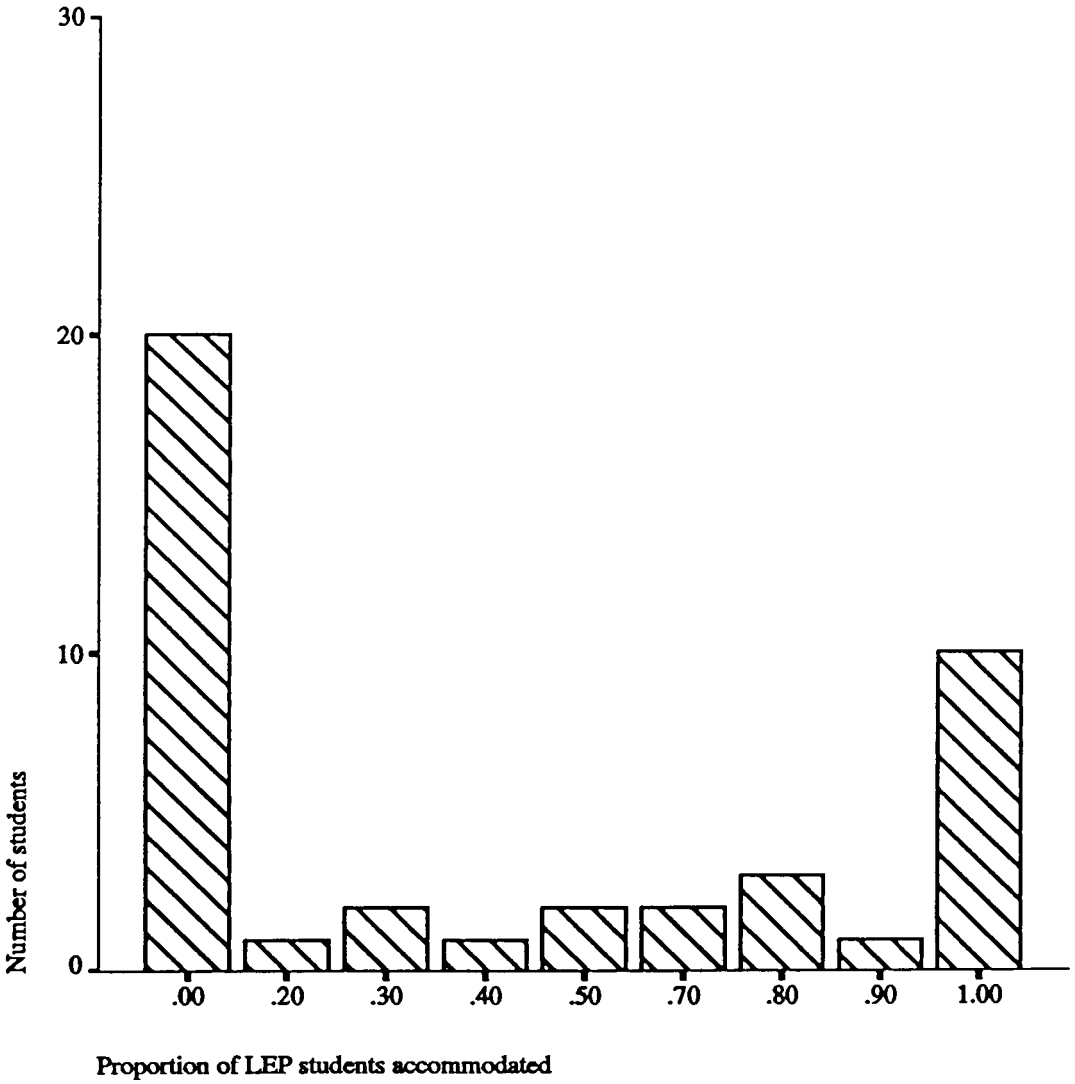

Figure 8. Number of schools, with fewer than ten LEP students, accommodating various proportions of LEP students ($n = 42$).

special education students. For schools with higher numbers of special education students, the use of accommodations was more broadly distributed and showed less evidence of extreme practices.

To see if all-or-none practices reflected schoolwide policies or attitudes regarding accommodations, we also calculated correlations between the two proportions. Was there a relationship between the proportion of LEP students in a school accommodated and the proportion of special education students accommodated? Overall this correlation was only .38 based on a total of 68

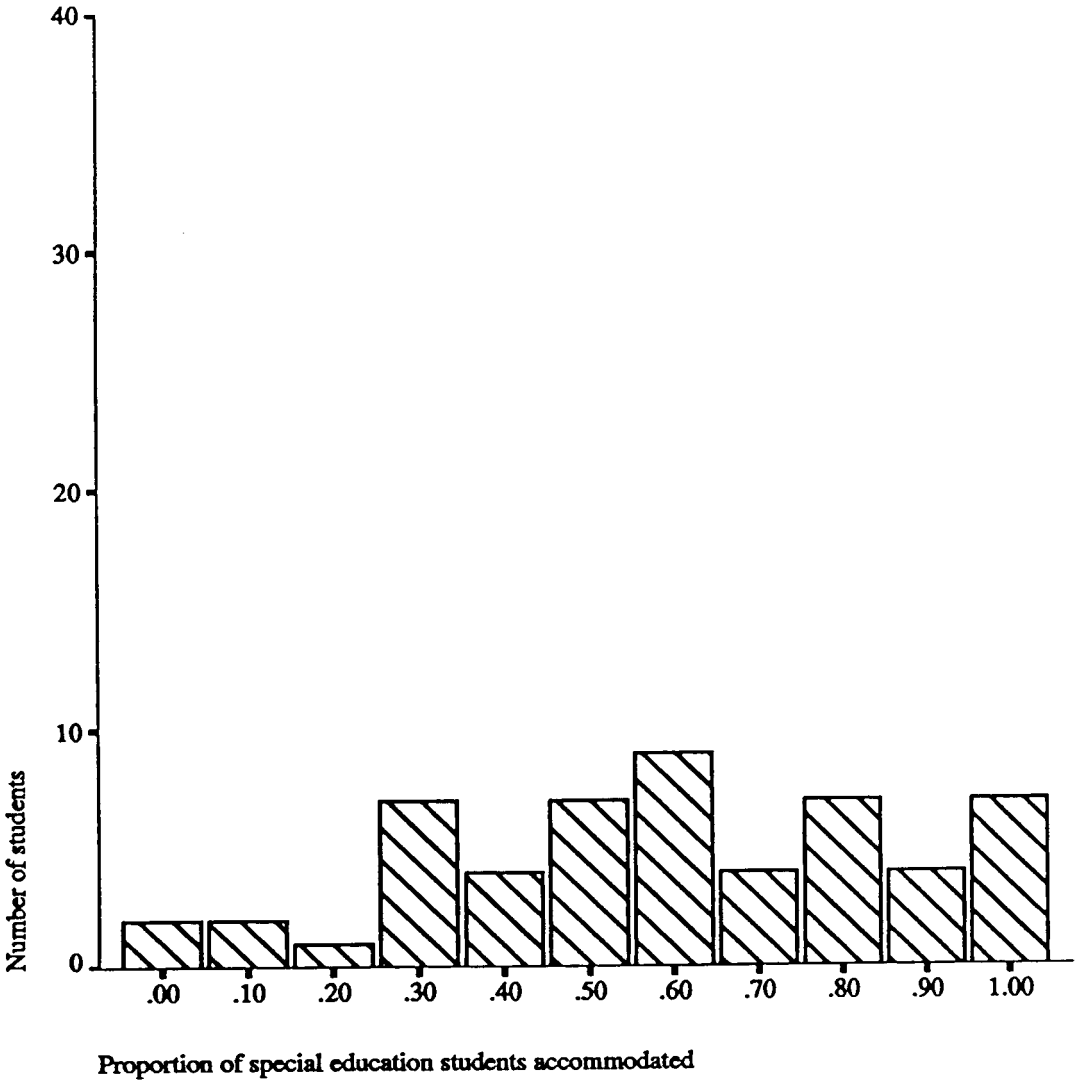

Figure 9. Number of schools, with ten or more special education students, accommodating various proportions of special education students (n = 54).

schools where both populations were represented. However, the correlation was near zero (r = .08) between the accommodation rates for the more severely affected groups, LEP students with less than 2 years in the U.S. and special education students in separate placements more than 50% of the school day. We reasoned that accommodation decisions for these two groups would logically be unrelated because they would most likely be made by different teachers in their respective self-contained settings. In contrast, the school-level relationship, between the proportion of accommodations provided for LEP students with

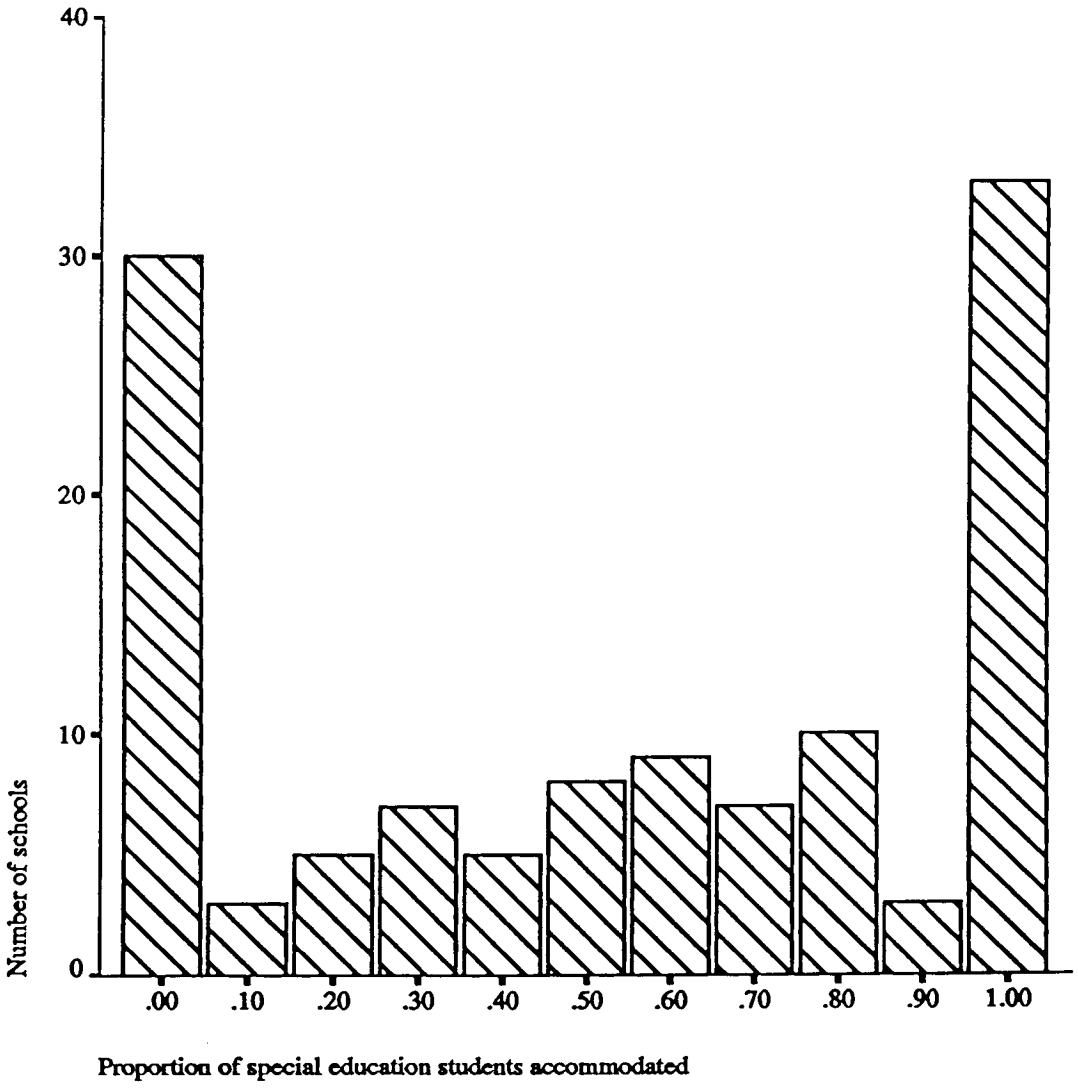

Figure 10. Number of schools, with fewer than ten special education students, accommodating various proportions of special education students (n = 120).

more than 2 years of education in the U.S. and the proportion of more mildly affected special education students who received accommodations, was .53. This suggests that for more mildly affected groups, where classroom teachers would be making the accommodation decisions, there were schoolwide tendencies to accommodate all or to accommodate none. Although many of these instances might indeed be justified, as an overall pattern it raises questions about the validity of accommodation decisions.

Performance Results for Students With Specific Accommodations

As shown previously in Table 2, the statewide mean on the Mathematics Performance Assessment for general education students was 15.91. Whether they received accommodations or not, LEP and special education students scored consistently below this level but not as far below as on the Metropolitan Achievement Test. Data in Table 3 show the frequency with which various accommodations were used and the associated performance levels for each. Many of the available accommodations such as oral administration of the test in Spanish, use of translation dictionaries, or giving a response orally in Spanish are omitted from the table because they occurred with such low frequency.

The most widely used accommodations involved changes in administrative procedures, especially oral reading of the assessment, repeating directions, testing in a special classroom or small group, and extended time limits. Some of these adaptations were so popular that they were applied to general education students as well. In fact, there were almost as many general education students who received extended time (n = 424) as there were LEP and special education students who received this accommodation (total n = 436).[2]

Assessment results necessarily confound the effect of the accommodation with initial differences in examinees' language proficiencies, which led to the accommodation decision. For example, the LEP (≥ 2 years in the U.S.) students who received the "repeating directions" accommodation scored relatively well ($\bar{X}$ = 10.87), although still well below the state average. Presumably these students were selected for this very limited accommodation because they were expected to function well with the regular assessment. It is not known, however, how these students would have fared if they had also had some kind of language support. Generally, it is expected that there will be a correlation between students' proficiency in English and their level of academic achievement, unless they are recent immigrants with strong academic preparation in their native language. LEP students who received translated versions of the assessment performed very poorly. A possible explanation is that these students are not receiving sufficient instruction in mathematics to be able to do the level of

[2] The state allows an additional 10 minutes each day if students are not finished with the assessment. This provision applies to all students. The accommodation of "extended time" means that students were given extra time beyond the usual 10 minutes. Perhaps some teachers are mistakenly reporting the allowable 10 minutes as an "extended time" accommodation.

Table 3

Statewide Means for Fourth Graders Who Received Specific Accommodations on the Mathematics Performance Assessment

		General education	LEP < 2 years	LEP ≥ 2 years	Special education ≥ 50%	Special education < 50%
	Administrative accommodations					
04	Oral Reading of assessment	12.90 ($n = 154$)	9.07 ($n = 29$)	9.37 ($n = 193$)	7.49 ($n = 259$)	12.51 ($n = 263$)
06	Repeating directions	12.25 ($n = 178$)	7.96 ($n = 25$)	10.87 ($n = 166$)	7.74 ($n = 222$)	12.79 ($n = 281$)
08	Written translation of assessment into Spanish	2.00 ($n = 3$)	5.56 ($n = 36$)	5.55 ($n = 22$)		5.00 ($n = 5$)
09	Oral translation of assessment into Spanish	2.00 ($n = 3$)	7.77 ($n = 26$)	7.37 ($n = 19$)		6.43 ($n = 7$)
	Response accommodations					
23	Giving response orally (written verbatim by test administrator)	19.30 ($n = 10$)	15.50 ($n = 4$)	10.00 ($n = 16$)	8.05 ($n = 57$)	14.41 ($n = 34$)
26	Writing response in Spanish	5.00 ($n = 1$)	6.33 ($n = 15$)	5.65 ($n = 20$)		5.00 ($n = 4$)
38	Adult transcription of portion of student's writing	13.71 ($n = 7$)		17.33 ($n = 3$)	7.74 ($n = 19$)	15.81 ($n = 16$)
	Setting accommodations					
40	Testing in special education or resource room	12.14 ($n = 28$)	8.00 ($n = 1$)	11.20 ($n = 10$)	8.13 ($n = 201$)	13.80 ($n = 250$)
41	Testing with small group	13.83 ($n = 93$)	8.16 ($n = 19$)	8.64 ($n = 44$)	8.72 ($n = 146$)	13.64 ($n = 215$)
42	Testing individually	17.20 ($n = 10$)	8.50 ($n = 2$)	9.67 ($n = 27$)	8.02 ($n = 57$)	13.97 ($n = 32$)
43	Testing with student seated in front of classroom	13.38 ($n = 13$)	2.00 ($n = 2$)	1.50 ($n = 2$)	5.75 ($n = 4$)	11.56 ($n = 9$)
45	Testing in ESL classroom	6.40 ($n = 15$)	9.00 ($n = 37$)	11.44 ($n = 204$)	5.00 ($n = 2$)	11.13 ($n = 15$)
50	Extended time	17.15 ($n = 424$)	8.71 ($n = 24$)	9.89 ($n = 138$)	10.48 ($n = 130$)	14.00 ($n = 144$)
51	More frequent breaks during testing	10.13 ($n = 16$)	6.50 ($n = 2$)	9.55 ($n = 11$)	8.44 ($n = 89$)	14.93 ($n = 40$)
52	Extended testing sessions over several days	16.29 ($n = 7$)	13.67 ($n = 3$)	6.64 ($n = 25$)	7.62 ($n = 29$)	15.07 ($n = 14$)

work required on the 4th-grade assessment. But it is also possible, that the Spanish version of the assessment is not comprehensible to them. This would occur, for example, if students are not fully literate in Spanish but are administered the written translation. A third possibility is that the Spanish and English versions of the assessments were not equated adequately. Although in-depth validity studies would be required to sort out these effects in a definitive way, it would also be useful simply to ask teachers why they believe that performance is so low for students in these groups.

Table 4 was constructed in an attempt to disentangle selection effects—that is, low-achieving students being selected to receive accommodations—from effects of the accommodations themselves. For each of the most frequently used accommodations, data in Table 4 report the number of accommodated students who took both the performance assessment and the Metropolitan Achievement Test, the Mathematics Performance Assessment mean for these students, and a standardized "improvement" score indicating the relative gain on the performance assessment compared to the MAT in standard deviation units.

One hypothesis was that the availability of accommodations would increase the number of low-achieving students who participated in the assessment. By comparing the sample sizes and the performance assessment means in Tables 3 and 4, it is possible to see whether the greater number of students who took only the performance assessment lowered the performance level compared to the means for students who took both tests. Indeed, in many instances the means are slightly higher in Table 4 than in Table 3. For example, reading across the tables for the oral reading accommodation provided to various groups, the means were 12.90 vs. 13.13, 9.07 vs. 8.33, 9.37 vs. 9.91, 7.49 vs. 7.90, and 12.51 vs. 12.54. In four of the five comparisons the means were higher for the more select group that took both tests (Table 4). However, in general, these differences were surprisingly small. Lack of substantial selection effects might be due to school-to-school differences in the decision to exclude students from taking the MAT, just as we observed tremendous differences among schools in the use of accommodations.

For those students who took both the MAT and an accommodated version of the performance assessment, it was possible to document the relative gain or improvement in performance associated with the accommodation. For example, in Table 4, the accommodation of orally reading the assessment to students

Table 4

Means on the Mathematics Performance Assessment, and Relative Z-Score Improvement on the Performance Assessment Compared to the Metropolitan Achievement Test, for LEP and Special Education Students Receiving the Most Frequently Used Accommodations

Accommodation	General education	LEP < 2 years	LEP ≥ 2 years	Special education ≥ 50%	Special education < 50%
04 Oral reading of assessment	13.13 .49 (n = 119)	8.33 .04 (n = 12)	9.91 .49 (n = 127)	7.90 .60 (n = 91)	12.54 .87 (n = 212)
06 Repeating directions	12.30 .35 (n = 149)	8.13 -.09 (n = 8)	11.47 .50 (n = 112)	8.27 .56 (n = 82)	12.79 .38 (n = 231)
08 Written translation of assessment into Spanish		7.58 .26 (n = 12)	4.50 .08 (n = 8)		3.50 .06 (n = 2)
40 Testing in special education or resource room	13.21 .50 (n = 19)			7.90 .53 (n = 61)	14.10 .42 (n = 208)
41 Testing with small group	13.87 .46 (n = 79)	10.00 -.02 (n = 7)	7.79 .03 (n = 29)	8.13 .56 (n = 68)	13.65 .34 (n = 181)
45 Testing in ESL classroom	7.33 .06 (n = 3)	9.63 .49 (n = 16)	11.74 .60 (n = 154)		11.00 .92 (n = 11)
50 Extended time	17.34 .39 (n = 371)	9.50 .38 (n = 4)	9.67 .38 (n = 101)	9.22 .73 (n = 55)	14.20 .72 (n = 122)

improved the performance of LEP students with 2 or more years in the U.S. by half a standard deviation (z = .49) compared to the performance of these same students on the MAT. It is not appropriate to try to interpret results for LEP students with less than 2 years of education in the U.S. because the sample sizes are so small. Statewide, a total of only 139 students in this category participated in the performance assessment, 94 of whom received accommodations. The numbers are very small for specific accommodations and for matched data on the MAT. For LEP students with more than 2 years education in the U.S., Spanish language accommodations also occurred too infrequently to be

interpreted meaningfully but several of the administrative accommodations had substantial effects, improving performance by .49, .50, .60, and .38 standard deviations. Administrative accommodations also had consistently large effects on the performance of special education students, improving performance compared to how the same students did on the MAT by .34 to .92 standard deviations.

Relative Performance on the Metropolitan Achievement Test and the Rhode Island Mathematics Performance Assessment

When comparing the results for the Metropolitan Achievement Test and the Rhode Island Performance Assessment in mathematics, it has already been shown that both LEP students and special education students performed relatively better on the performance assessment. In the analysis accompanying Table 2, we focused on the respective statewide populations. On average, LEP and special education students were not as far below general education students on the performance assessment, even though greater inclusion on the performance assessment had most likely increased the number of lower achieving students who participated.

In the analyses that follow, we use only the matched data sets of students who took both the MAT and the performance assessment. This allows us to "control" any selection biases due to differences in participation on the two tests. It also allows comparison of relative differences in performance across the entire distribution of scores rather than comparing only mean score differences. In Figures 11 and 12, major-axis plots are shown for general education students with data for LEP students and special education students superimposed. These graphs illustrate the strong correlation between the two different measures of mathematics achievement (r = .73). Major-axis lines of best fit differ from more familiar regression lines by minimizing errors on both the x and y dimensions simultaneously. A regression line helps answer the question, what is the most likely score on y, given a score on x. But a second regression line is needed to describe the relationship, if, instead, y is used to predict x. The major-axis line of best fit is a symmetrical solution that defines, on average, the equating of one variable with the other. When both variables are reported as standardized z-scores, the major-axis is the 45 degree line.

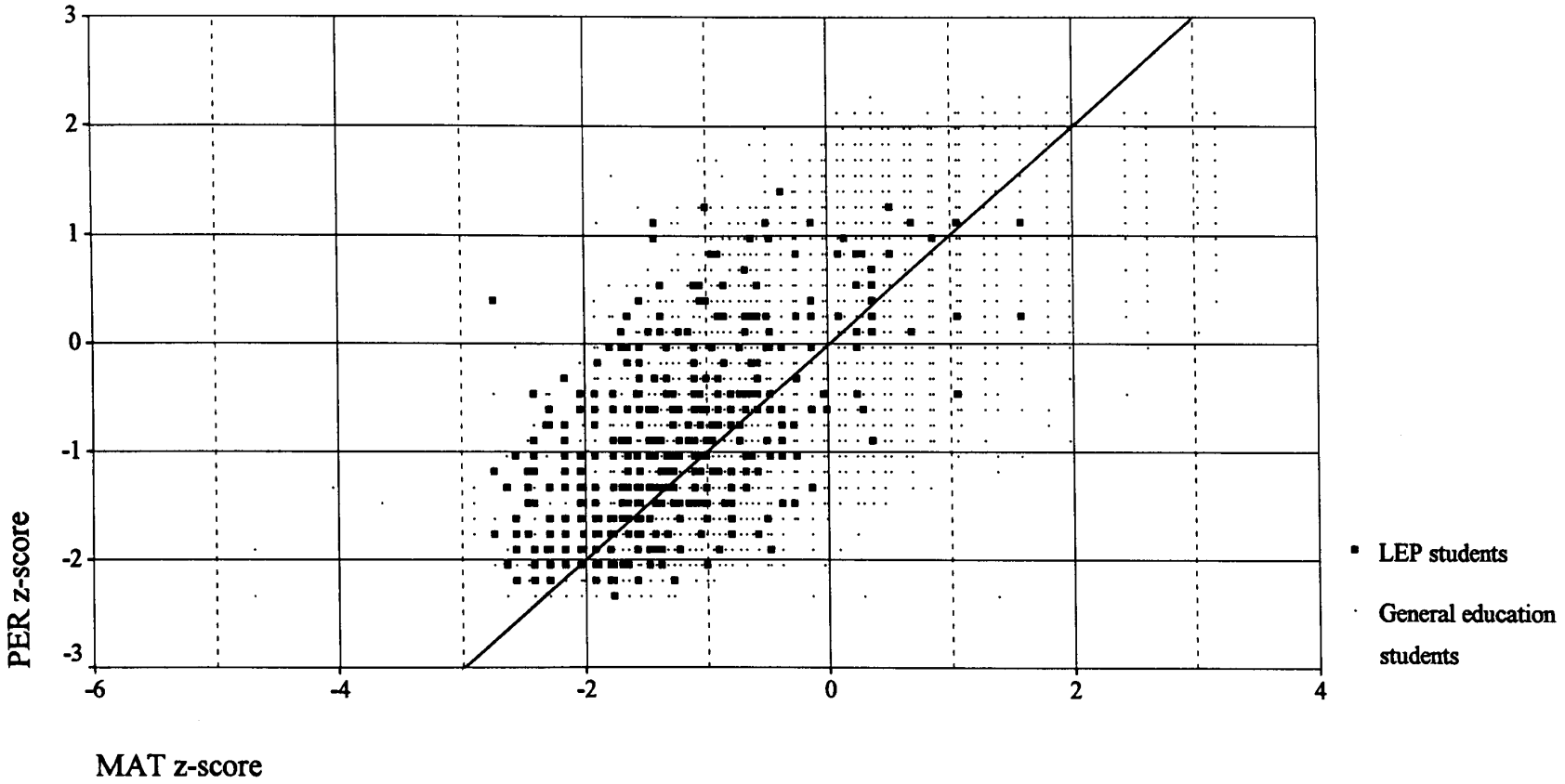

Figure 11. Scatterplot depicting major axis for general education students with data for LEP students superimposed.

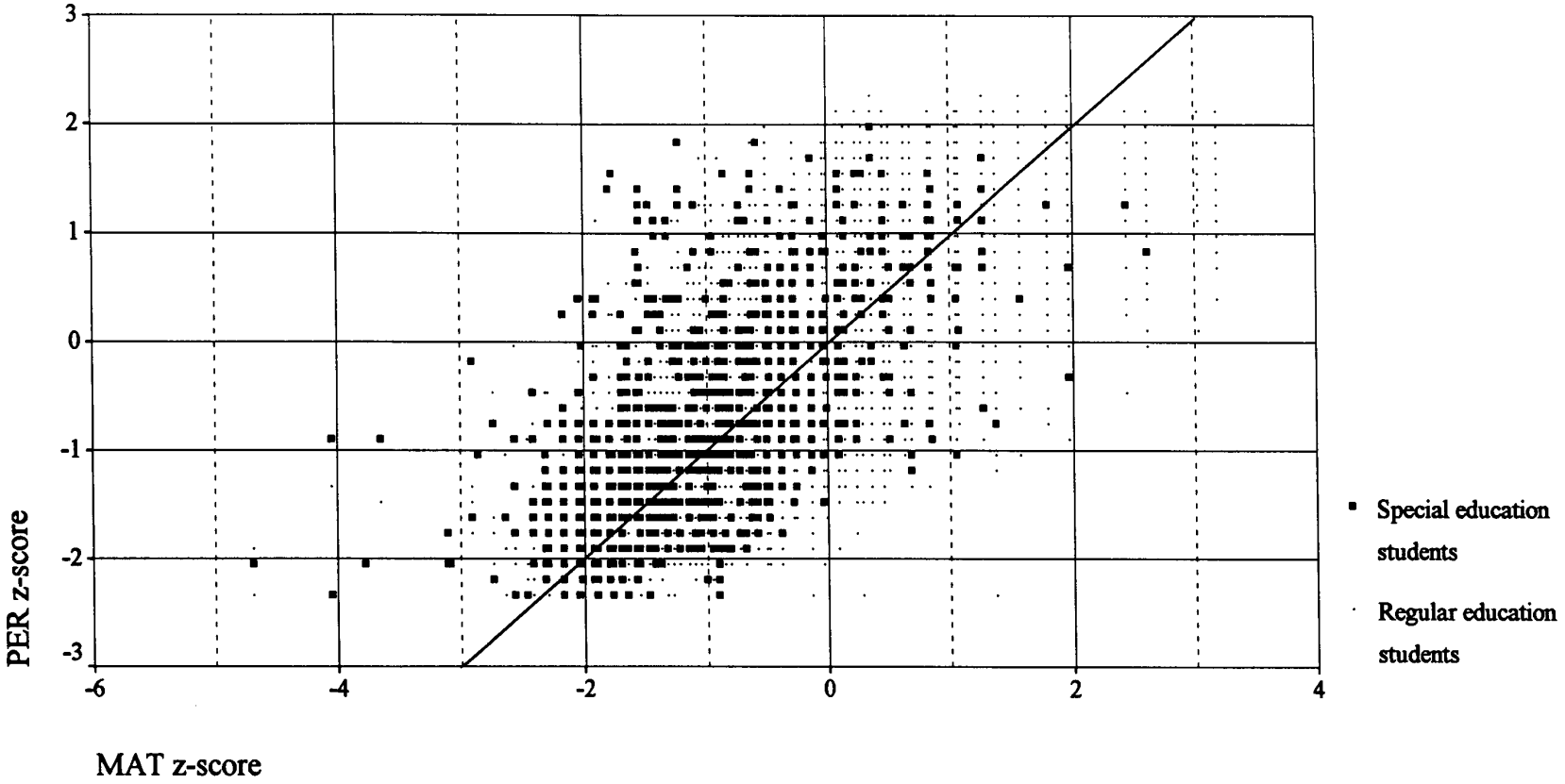

Figure 12. Scatterplot depicting major axis for general education students with data for special education students superimposed.

In Figure 11, the relative advantage of LEP students on the Mathematics Performance Assessment compared to the MAT can be seen as the greater density of LEP data points above the major-axis line. The magnitude of the advantage can also be evaluated by visual inspection. Note, for example, how many LEP students scored more than 2 standard deviations below the mean on the MAT (z = -2), while a much smaller number had performance assessment scores below a -2. The same relative advantage is also apparent when comparing z scores of -1, 0, and so forth. A similar pattern of relative advantage is also apparent in Figure 12 for special education students. In fact, the effect appears to be even more pronounced for special education students because the dispersion of scores far above the major-axis equating line is greater and because there are more special education students at the higher achievement levels on both tests.

Did LEP and special education students do better on the performance assessment because they received accommodations? Data in Table 5 show the relative gain on the performance assessment compared to the MAT in standardized units for accommodated versus non-accommodated LEP and special education students. Clearly, the relative advantage on the performance assessment is much greater for accommodated students.

Are the performance gains caused by the use of accommodations valid? Of course, this question cannot be answered without additional criterion validity data. Accommodations should improve performance by allowing students a

Table 5

Relative Z-Score Improvement on the Performance Assessment Compared to the Metropolitan Achievement Test, for LEP and Special Education Students Who Were and Were Not Accommodated ($ES_{PA} - ES_{MAT}$)

	LEP < 2 years	LEP ≥ 2 years	Special education ≥ 50%	Special education < 50%
Accommodated	.32 (n = 30)	.51 (n = 289)	.50 (n = 127)	.42 (n = 388)
Not accommodated	.02 (n = 13)	.10 (n = 132)	.06 (n = 46)	.08 (n = 394)
Total	.22 (n = 43)	.38 (n = 421)	.38 (n = 173)	.25 (n = 782)

better opportunity to demonstrate their true level of learning. Some of the data in Figures 11 and 12, however, raise a question about whether accommodations were used appropriately. In some cases, seen at the top edge of the scatter-plot ellipse, the gains from the MAT to the performance assessments were so great—1 or 2 standard deviations—as to raise questions about their credibility. Out of the 464 LEP students statewide who took both tests, 319 students were accommodated, and of these 111 had relative gains on the performance assessment of .70 standard deviations or more. Of these more remarkably large gains that raise questions about the validity of accommodations, 64 occurred in only four schools. The more typical pattern is for schools to have no LEP students who gained such substantial amounts from accommodation or only one or two such students. Therefore, it is reasonable to call into question the practices of schools where 13 to 24 LEP students made huge gains. The improvement caused by accommodations was so great in these four schools that even when the flagging cut point was doubled (from .7 to 1.4), there were still 3, 4, 4, and 13 LEP students with relative advantages greater than this amount. Large-scale assessment programs may wish to add a statistical flagging procedure such as this to check on the appropriateness of accommodation practices; however, it is possible to detect these extreme shifts only because of the availability of MAT data on some students.

Data From the Pilot Sample

The purpose of the pilot study was to gather collateral data, in addition to the state-administered standardized test and performance assessment, that would provide preliminary evidence about how the two measures functioned for language-minority students compared to monolingual English speakers. Data were collected for 443 students from 22 volunteer classrooms selected from schools with relatively higher concentrations of language-minority students. Table 6 shows the distribution of different levels of language proficiency for the total sample and for subgroups of students who participated in the performance assessment, the MAT, or both. Despite the presence of a relatively large population of English-language learners in these schools, representing one third of the entire sample, the sampling procedure did not yield sufficient numbers of LEP students, especially because it would be desirable to analyze accommodated students separate from non-accommodated students. Most 4th-graders were considered advanced or advanced intermediate English learners. Of the small

Table 6

Participation of Pilot Sample Students in the Mathematics Performance Assessment and Metropolitan Achievement Test

Teacher language proficiency ratings	Total sample	Participated in PA	Participated in MAT	Participated in both tests	Identified as LEP on the PA
06 Monolingual English	294	220	281	217	0
05 Nonnative Advanced	95	72	89	70	1
04 Advanced Intermediate	33	24	28	24	16
03 Intermediate	12	8	11	8	4
02 Advanced Beginner	6	2	3	2	2
01 Beginner	3	1	2	1	0
TOTAL	443	327	414	322	23

number of students, 21, with more limited English proficiency (categories 1, 2, 3), only 11 students participated in the performance assessment. Unlike the statewide results where participation rates were greater on the performance assessment, in the pilot sample classrooms participation was better on the MAT even for English-language learners.

In order to make the pilot analyses as parallel to the statewide analyses as possible, we preferred to use the LEP identification on the performance assessment rather than the language proficiency rating provided by the classroom teachers. However, use of the assessment-based designation led to a further loss of English-language learners from the analysis because their teachers had not coded them as LEP on the formal assessment. At the same time, the assessment-based LEP classification included additional students whom teachers had rated as advanced intermediate (16) or advanced (1) English speakers. In this study, inconsistent labeling of LEP students caused serious problems with attrition, and consequently with sample size, but it also tells us that even in the statewide study misclassification of students can confound the evaluation of comparisons between general education and LEP students. To increase the

numbers of LEP students in the pilot sample analysis as much as possible, but still be certain that students' language skills were limited, the final decision was to include LEP students identified on the performance assessment (23) plus students who received the lowest three ratings of language proficiency but were not identified as LEP on the performance assessment (an additional 5 students).

Achievement data for students in the pilot sample are summarized in Table 7. Means on the Mathematics Performance Assessment and the MAT are slightly below the state averages in Table 2. Nonetheless the pattern of results for subgroups is very similar between the statewide and pilot sample data. Table 7 also includes the category of advanced non-native English speaker, which is not found in the state-level data. In fact, once English-language learners move past the "limited" designation, they become indistinguishable from monolingual English speakers despite the fact that less than perfect English fluency may

Table 7

Means and Standard Deviations for Fourth Graders in the Pilot Sample on the Mathematics Performance Assessment and Metropolitan Achievement Test

	Monolingual general education	Monolingual special education	Advanced nonnative speaker	LEP
Performance Assessment (PA)	15.48 6.66 (n = 200)	11.63 6.47 (n = 19)	12.08 7.20 (n = 72)	11.25 5.82 (n = 28)
PA, matched data set	15.48 6.66 (n = 200)	9.94 5.08 (n = 16)	12.13 7.21 (n = 70)	11.25 5.82 (n = 28)
Metropolitan Achievement Test (MAT)	604.28 41.26 (n =254)	562.75 33.05 (n = 16)	583.49 35.86 (n = 89)	560.67 .37.90 (n = 33)
MAT, matched data set	609.11 41.79 (n = 200)	562.75 33.05 (n = 16)	584.66 35.86 (n = 70)	566.29 25.25 (n = 28)
Average math grade	3.02 .89 (n = 262)	2.41 .52 (n = 19)	2.85 .95 (n = 90)	2.53 .99 (n = 36)
Standards-based rating	2.54 .99 (n = 264)	1.63 .60 (n = 19)	2.38 .96 (n = 95)	1.84 .82 (n = 38)

continue to affect their academic performance (Cummins, 1979). Advanced non-native English speakers achieve at a much higher level than LEP students on both the performance assessment and the MAT but are still substantially below the general education averages. Data provided by classroom teachers are also shown for both average mathematics grade (first semester and third-quarter combined) and a standards-based rating. Special education and LEP students were further behind the other groups on the standards-based rating than on the math grades, which would be expected if standards represented a common and absolute scale but grades were adjusted to reflect expectations for the group or individualized education plans.

Table 8 shows the use of specific accommodations for LEP students in the pilot sample. Performance results are not reported because they could be misleading with such small numbers. Of the 25 students identified as LEP on the performance assessment, only 12 received one or more accommodations. As was the case for the entire state, administrative accommodations were the most frequent, especially oral reading of the test, repeating directions, testing in the ESL classroom, and providing extended time. Of the 12 students accommodated, 9 received two or more accommodations, with the most frequent pattern being oral reading and repeating directions in the ESL classroom.

Table 8

Numbers of LEP Students in the Pilot Sample Who Received Specific Accommodations on the Mathematics Performance Assessment

	LEP
04 Oral reading of test	9
05 Signing of assessment	1
06 Repeated directions	11
08 Written translation into Spanish	1
38 Adult transcription of portion of student's work	1
40 Testing in special education class	1
41 Testing with small group	7
42 Testing individually	3
45 Testing in ESL class	8
50 Extended time	8

Relative Performance on the Metropolitan Achievement Test and the Rhode Island Mathematics Performance Assessment in the Pilot Sample

In the statewide analyses the performance assessment and MAT correlated .73. In the pilot sample the correlation was .74. The relationship between the two measures is depicted in Figure 13. Again the major-axis equating line was established for general education students. As was true in the state analysis, LEP students in the pilot sample did relatively better on the performance assessment gaining .24 standard deviation units compared to their average score on the MAT. This effect size was not quite so large as for the state as a whole, where the relative gain was .36 standard deviation units, probably because a smaller proportion of LEP students were accommodated in the pilot sample. The relative performance of special education students compared to general education students is shown in Figure 14. In this case, the results were different from the statewide results, with no overall advantage for special education students on the performance assessment. However, for the lowest achieving special education students ($z \leq -1$), there was a definite advantage on the performance assessment.

Because the relative advantage for LEP students on the performance assessment was not as great in the pilot sample as for the state as a whole, there was less reason to question the validity of the performance assessment results. For example, if we chose to scrutinize those scores more closely where LEP students scored a standard deviation (or more) higher on the performance assessment than on the MAT, we would identify five LEP students, or 19% of the LEP sample. However, the same proportion of general education students had difference scores of similar magnitude. With this caveat in mind, we examined some of the outlier cases merely to illustrate how collateral data might be used. Keep in mind, however, that there was no systematic evidence in the pilot sample that schools might be misusing accommodations.

Case 1 can be seen in Figure 13 as the LEP student with the highest performance assessment score ($z = 1.5$) but with a MAT score almost one standard deviation below the mean. This student received two accommodations on the mathematics performance assessment: She was tested individually, in the ESL classroom. According to her classroom teacher, her grades in mathematics were at a C level throughout the year and her standards-based rating in mathematics was Below Basic. Case 1 is clearly an outlier in the major-axis plot

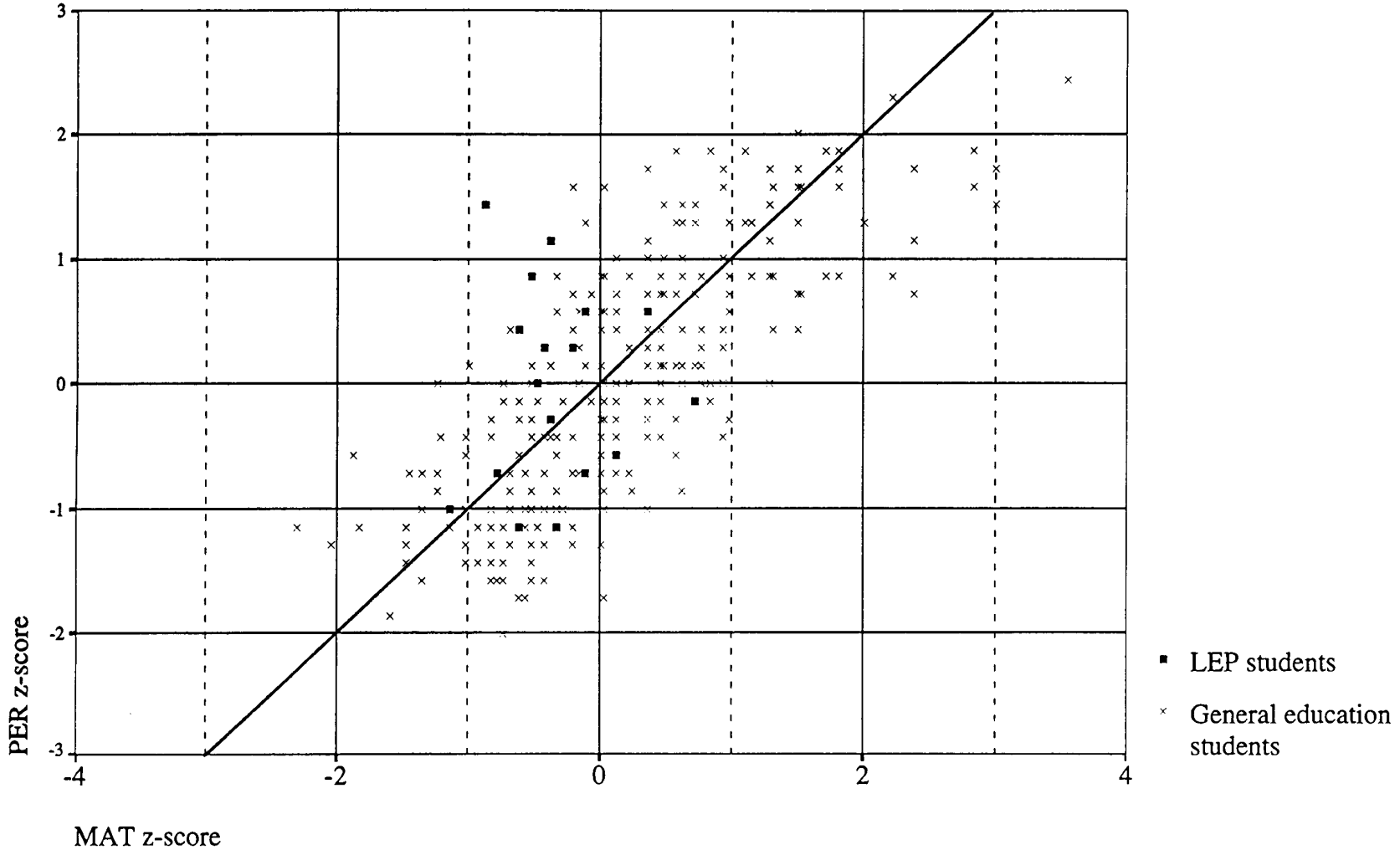

Figure 13. Scatterplot depicting major axis for general education students with data for LEP students superimposed (Pilot sample).

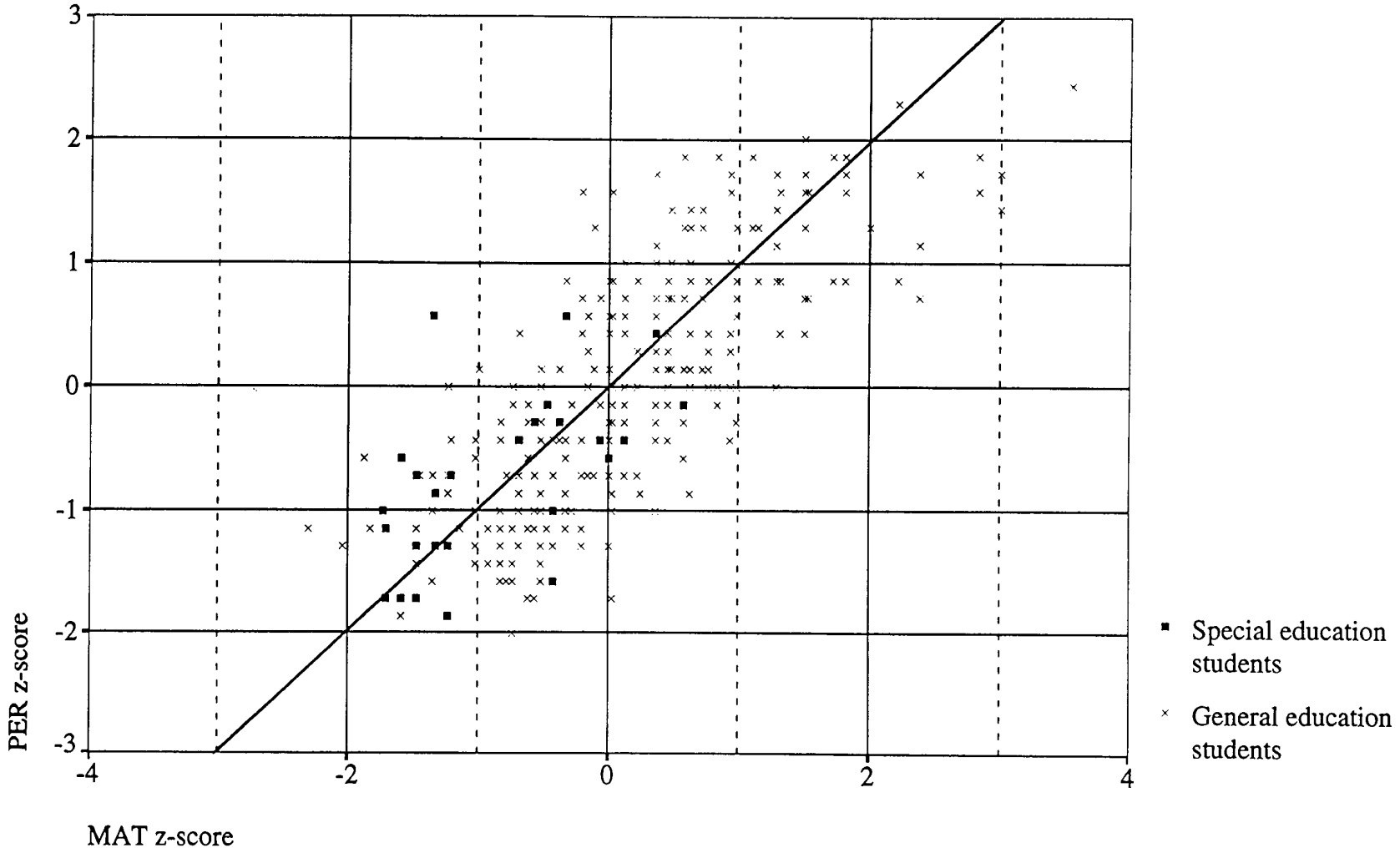

Figure 14. Scatterplot depicting major axis for general education students with data for special education students superimposed (Pilot sample).

and is most likely an example of an invalidly inflated score resulting from accommodation.

Case 2 occurred in the same school as Case 1 and can be seen in Figure 13 as the second highest LEP score on the performance assessment, while still being below the mean on the MAT. In the statewide analysis, Case 1 and Case 2 were the only two extreme gains associated with accommodations in that school, despite there being other LEP students, suggesting that there was not a pervasive misuse of accommodations there. Case 2 was also tested individually and received the accommodation of oral reading of the assessment. In this example, however, the student received mathematics grades of 3.0 and 3.33 and a standards-based rating from her classroom teacher of Basic. Case 2 also happened to be one of the cases selected by the classroom teacher to represent students in the third quartile of the class (next-to-bottom quartile) and for whom student work was collected. The work samples were quite consistent with the teacher's Basic rating and reflected excellent computational skills. For example, the student answered flawlessly fill-in-the-box equation problems involving addition, subtraction, and multiplication. She made almost no errors on worksheets requiring the use of quantitative information from charts and answered most simple word problems correctly. For example, "Linda had 4 quarters, 5 dimes, and 2 nickels. She gave 4 quarters and 4 dimes to her brother. What coins does Linda have left?" Case 2 could not complete pattern problems and showed a lack of understanding of how different areas on a spinner problem would affect the outcome of a game. Our conclusion, after comparing the student's classroom work with her below average MAT score and substantially above average performance score, was that the truth was probably somewhere in between. In fact, we concur with her teacher's rating of Basic, which means that the MAT underestimated her true mathematics proficiency and the accommodated performance assessment overestimated it.

Case 3 received the third highest performance assessment score for LEP students in Figure 3 and a significantly below average MAT score. This student did not receive an accommodation despite a language rating of intermediate. She had grades in mathematics throughout the year of 4.0 but a standards-based rating of Basic, which probably means that her A grades were in relation to an individualized standard. Classroom work was also collected for Case 3, which consisted entirely of Silver Burdett worksheets that closely resembled problems

on the MAT. Case 3 could do many of the computational and word problems including those involving fractions, but she often missed items involving division or more difficult multiplication. Her weekly "tests" showed an average of from 55% to 81% correct. Again we agreed with her teacher's rating of Basic, suggesting an inflated result on the performance assessment but in this case it cannot be attributed to the use of an accommodation.

These cases were specifically selected as instances of extreme discrepancies between the two tests. Therefore, they do not reflect a generalized problem of inflated performance assessment results for LEP students. They do suggest that *some* accommodated scores may be inflated and may therefore detract from rather than enhance the validity of assessment results. Note that we did not examine cases where discrepancy scores favored LEP students on the MAT because there were zero students with a relative *z* score advantage on the MAT greater than 1.

Validity Correlations: Assessment Data and Teachers' Ratings

In addition to the simple correlation between the Mathematics Performance Assessment and the Metropolitan Achievement Test, the validity of both measures can be evaluated in comparison to classroom teachers' ratings of students' proficiency in the pilot group. Of particular interest is whether the degree of validity correlations among these variables found for monolingual English test-takers holds true for language-minority students.

Correlations are reported in Table 9 for monolingual general education students in the pilot sample. Although teachers provided data on 267 students in this category, matched data on the two tests were available for only 193 students because of non- participation in testing, especially on the performance assessment. The strongest correlation (r = .75) was between the performance assessment and the MAT. The next highest value was the correlation between teachers' mathematics grades and teachers' standards-based ratings (r = .67). Other correlations among the tests and the teacher variables were substantial, ranging from .53 to .58, but were not as high as the test-test or teacher-teacher correlations. This pattern is to be expected given that teachers were not trained to ensure consistency of ratings across classrooms.

Table 9

Correlations Between Tests and Teachers' Ratings for Monolingual General Education Students in the Pilot Sample

	Metropolitan Achievement Test (MAT)	Performance Assessment (PA)	Teachers' mathematics grades	Teachers' standards-based rating
Metropolitan Achievement Test (MAT)		.75* (n = 193)	.53* (n = 255)	.58* (n = 257)
Performance Assessment (PA)	.75* (n = 193)		.55* (n = 193)	.58* (n = 193)
Teachers' mathematics grades	.53* (n = 255)	.55* (n = 193)		.67* (n = 265)
Teachers' standards-based rating	.58* (n = 257)	.58* (n = 193)	.67* (n = 265)	

*Significant at $p < .01$ level.

Data in Table 10 are the correlations based on the 27 accommodated monolingual special education students in the pilot sample. Teachers' grades and standards-based ratings were again strongly related (r = .72). Several other correlations were also significantly not zero. Given the small sample size, it is not warranted to try to interpret differences in correlations; for example, the MAT-performance assessment correlation is weaker here (r = .65) than for general education students, a finding which if reliable might be attributable to range restriction or to a change in the relationship due to accommodations. To illustrate the kinds of insights that could be gained from these kinds of analyses with more data, we note that teachers' standards-based ratings were more highly correlated with the MAT than with the performance assessment. Could this mean that accommodations on the performance assessment reduced the validity of the assessment results? A more rigorously conducted study with a larger sample, but more importantly with careful training of teachers on the standards-based rating, would be needed to answer this question.

In the pilot sample classrooms, there were a total of 95 English-language learners who were rated by their teachers as advanced in their English

Table 10

Correlations Between Tests and Teachers' Ratings for Accommodated Special Education Students in the Pilot Sample

	Metropolitan Achievement Test (MAT)	Performance Assessment (PA)	Teachers' mathematics grades	Teachers' standards-based rating
Metropolitan Achievement Test (MAT)		.65* ($n = 24$)	.36 ($n = 24$)	.53* ($n = 24$)
Performance Assessment (PA)	.65* ($n = 24$)		.43* ($n = 27$)	.41* ($n = 27$)
Teachers' mathematics grades	.36* ($n = 24$)	.43* ($n = 27$)		.72* ($n = 27$)
Teachers' standards-based rating	.53* ($n = 24$)	.41* ($n = 27$)	.72* ($n = 27$)	

*Significant at $p < .01$ level.

proficiency. Fourteen of these students received accommodations on the performance assessment and are reported separately in Table 12. Correlational data for the remaining 81 are in Table 11; however, only 57 of these students participated in both the MAT and the performance assessment. The correlations in Table 11 closely parallel those reported for monolingual general education students in Table 9. This suggests that once students are proficient in English, both forms of assessment provide information that is equally accurate for language-minority students and monolingual English speakers. These relationships were also found in Table 12 for advanced nonnative speakers taking the MAT. However, in Table 12 the correlations for the performance assessment with the two teacher variables are lower, raising a question about whether accommodations could have attenuated the validity of the assessment for these students. Such an interpretation is less plausible, however, given the very high correlation ($r = .83$) between the performance assessment and the MAT. Given the small number of advanced non-native English speakers who received accommodations, it is best not to try to interpret a shift in the magnitude of correlations.

Table 11

Correlations Between Tests and Teachers' Ratings for Advanced Nonnative English Speakers in the Pilot Sample Who Received No Accommodations

	Metropolitan Achievement Test (MAT)	Performance Assessment (PA)	Teachers' mathematics grades	Teachers' standards-based rating
Metropolitan Achievement Test (MAT)		.76* ($n = 57$)	.52* ($n = 71$)	.62* ($n = 76$)
Performance Assessment (PA)	.76* ($n = 57$)		.52* ($n = 57$)	.49* ($n = 58$)
Teachers' mathematics grades	.52* ($n = 71$)	.52* ($n = 57$)		.64* ($n = 76$)
Teachers' standards-based rating	.62* ($n = 76$)	.49* ($n = 58$)	.64* ($n = 76$)	

*Significant at $p < .01$ level.

Table 12

Correlations Between Tests and Teachers' Ratings for Advanced Nonnative English Speakers in the Pilot Sample Who Received Accommodations

	Metropolitan Achievement Test (MAT)	Performance Assessment (PA)	Teachers' mathematics grades	Teachers' standards-based rating
Metropolitan Achievement Test (MAT)		.83* ($n = 13$)	.60* ($n = 13$)	.61* ($n = 13$)
Performance Assessment (PA)	.83* ($n = 13$)		.36 ($n = 14$)	.43 ($n = 14$)
Teachers' mathematics grades	.61* ($n = 13$)	.36 ($n = 14$)		.54* ($n = 14$)
Teachers' standards-based rating	.61* ($n = 13$)	.43 ($n = 14$)	.54* ($n = 14$)	

*Significant at $p < .01$ level.

The correlations in Table 13 show the relationships among the tests and teachers' ratings for LEP students in the pilot sample. It is an important commentary on the confounding of language learning, academic achievement, and measurement artifact that the strongest correlation was between teachers' ratings of language proficiency and scores on the MAT. There were also significant validity correlations between the MAT and performance assessment (r = .46) and between the tests and teachers' grades. These values suggest that the assessment results were not just random for LEP students, but they do not have the same level of accuracy for LEP students as for other groups of students. This is in part a range restriction problem, as illustrated in the major-axis plot in Figure 13. Weak correlations with validity criteria mean that a test is not very accurate in measuring differences in achievement *among* individuals in a group. Given the restricted range of achievement scores, however, it could still be reasonably accurate in locating those individuals within a certain range on the full performance continuum. Limited validity correlations also reflect genuine

Table 13

Correlations Between Tests and Teachers' Ratings for LEP Students in the Pilot Sample

	Metropolitan Achievement Test (MAT)	Performance Assessment (PA)	Teachers' mathematics grades	Teachers' standards-based rating	Teachers' language rating
Metropolitan Achievement Test (MAT)		.46* (n = 28)	.49* (n = 28)	.28* (n = 28)	.54* (n = 28)
Performance Assessment (PA)	.46* (n = 28)		.44* (n = 28)	.01 (n = 28)	.43* (n = 28)
Teachers' mathematics grades	.49* (n = 28)	.44* (n = 28)		.58* (n = 28)	-.01 (n = .28)
Teachers' standards-based rating	.28 (n = 28)	.01 (n = 28)	.58* (n = 28)		-.13 (n = 28)
Teachers' language rating	.54* (n = 28)	.43* (n = 28)	-.01 (n = 28)	-.13 (n = 28)	

*Significant at p < .01 level.

inaccuracies in measurement, where the test does not capture what students really know, as illustrated by some of the discrepancies between the MAT and performance assessment results. The lack of correlation with teachers' standards-based ratings for both the MAT and the performance assessment was primarily a range restriction problem given that all but four LEP students received ratings of Below Basic or Basic.

Keeping in mind the limitations of the pilot study, both with respect to sample size and limitations of the criterion measures themselves, the MAT and performance assessments appear to be functioning reasonably well for all groups except the LEP students. Strong correlations between the two measures and with teachers' classroom ratings hold true even for language-minority students who are proficient in English. The pattern of slightly weaker correlations between the performance assessment and criterion variables for groups of accommodated students suggests that *some* accommodated scores have reduced rather than increased validity. Nevertheless, there is still a consistent relationship between students' mathematics achievement in the classroom and their performance on the performance assessment with accommodations. For LEP students, the validity relationships are much weaker. This loss in accuracy is true for both the MAT, which was administered without accommodations, and the performance assessment with accommodations.

Other Insights From the Major-Axis Analyses

The major-axis analyses, intended to examine the relationship between the performance assessment and MAT, also provided some interesting insights about classroom-to-classroom differences. By definition, the major-axis describes the equating line where general education students do equally well on the two measures. As shown previously, LEP students do relatively better on the performance assessment. Relative strengths and weaknesses were also examined for each classroom. While most classrooms showed symmetrical results consistent with the overall pilot sample picture, a few classrooms showed dramatically different patterns where students did either much better on the performance assessment or much better on the MAT.

Figures 15 and 16 provide two such examples. Students in class 4 did remarkably better on the MAT. Although the sampling of student work was collected only for a short period, during that time all of the worksheets were from Macmillan/McGraw-Hill's *Mathematics in Action*. The items, all in

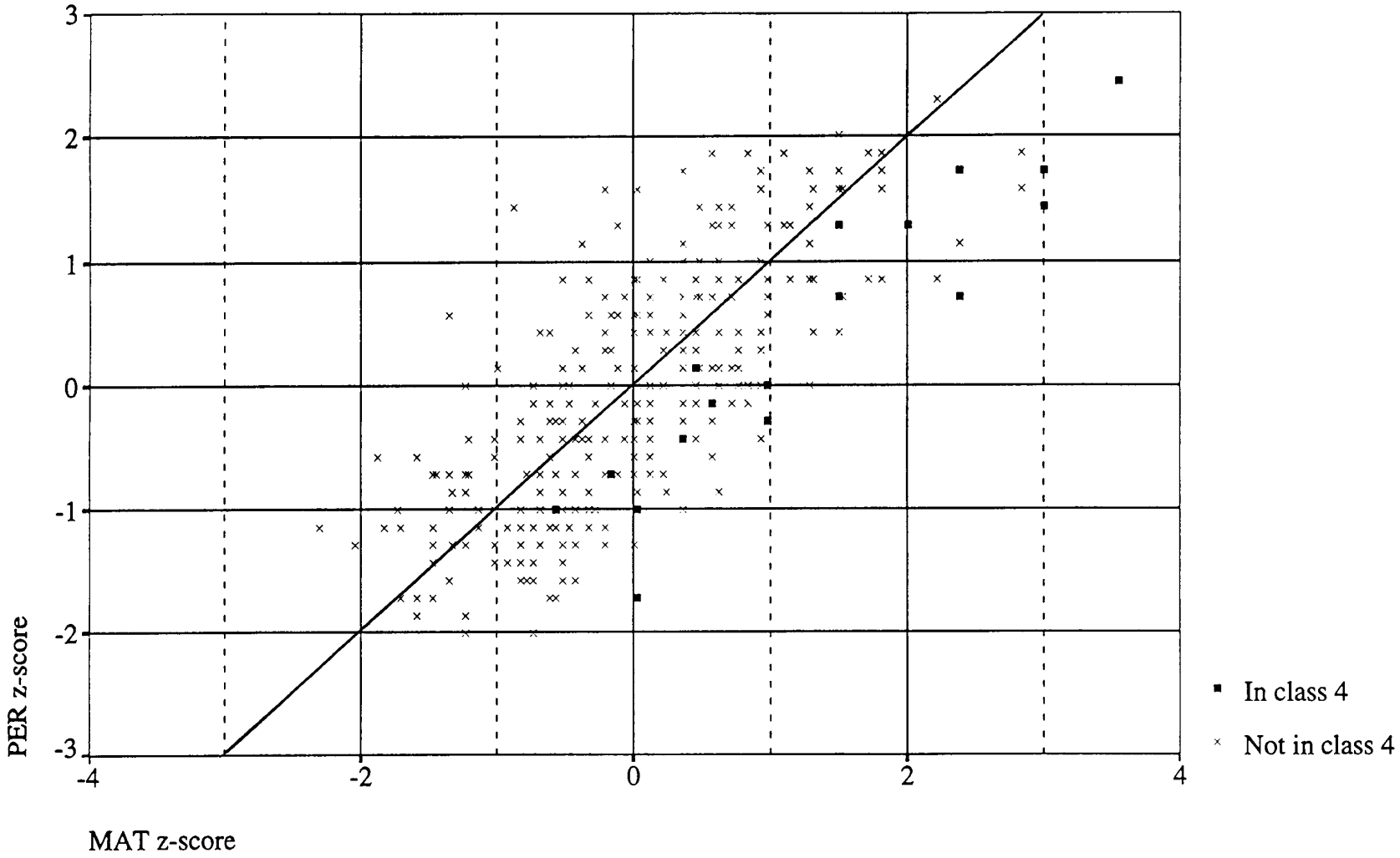

Figure 15. Scatterplot depicting major axis for students not in class 4 with data for class 4 students superimposed.

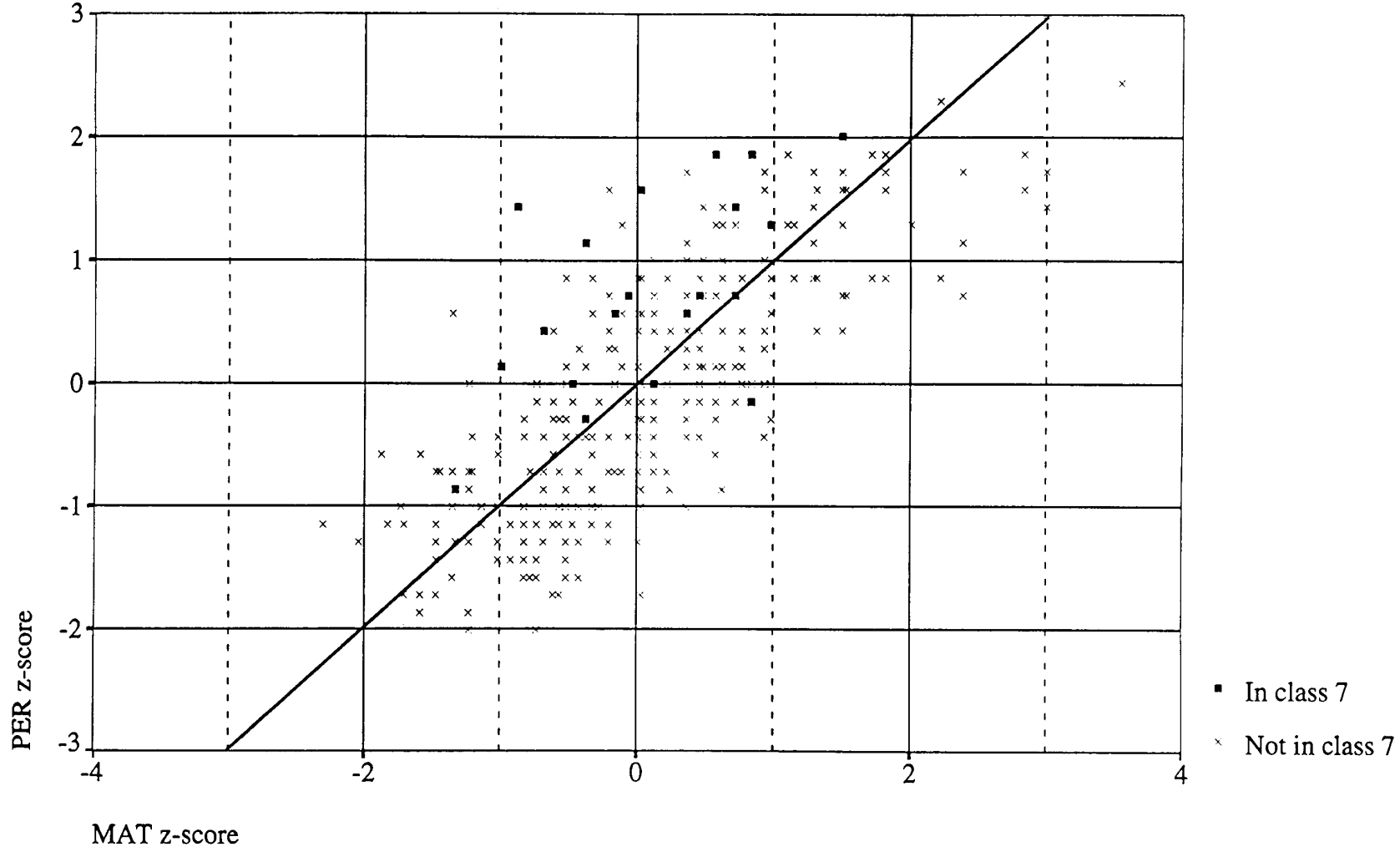

Figure 16. Scatterplot depicting major axis for students not in class 7 with data for class 7 students superimposed.

multiple-choice format, mapped closely to the types of items found on the MAT. Given that all but two of the 22 classes in the pilot sample used mathematics worksheets that resembled the MAT at least some of the time, it is surprising that classroom 4 stands out to such a marked degree.

Classroom 7 shown in Figure 16 is an example of the reverse effect. Here students did relatively much better on the performance assessment. When classroom work was examined for these students we found one of only two examples where students had been asked to construct answers to open-ended problems and where they had been asked to explain their work. Classroom 7 is the home of the Case 1 and Case 2 LEP students described above. This picture adds further evidence that, while their accommodated performance assessment scores might be somewhat inflated, their relative advantage on the performance assessment is credible and consistent with their classroom performance and the type of mathematics instruction that they and their classmates are receiving.

Differential Item Functioning on the Grade 4 Mathematics Performance Assessment

Differential item functioning (DIF) analyses were used to evaluate the relative difficulty of performance assessment items for LEP students. In previous analyses, it has already been shown that LEP students scored below the state average on the performance assessment. The purpose of the DIF analysis was to see whether the difference between the LEP and majority group is constant across all of the assessment tasks or whether there are some items that were relatively more difficult for LEP students. DIF statistics were at one time referred to as item-bias statistics because bias is one potential explanation for test items being relatively more difficult for one group than for other groups. However, there are other explanations for relative differences in item difficulty including differences in opportunity to learn.

Statewide data were used from the matched data set described in Table 1. Analyses were performed for the eight assessment tasks administered in common to all 4th graders in the state. Item response functions were estimated for 9,926 general and special education students and compared to those for 463 LEP students. Because each assessment task was scored on a 4-point scale, separate DIF statistics were calculated for each of the four score levels. See Appendix for further details on the methods of analysis.

Results of the DIF analyses are shown in Figure 17. When the value of the DIF statistic is .1 or lower, it means that LEP students are answering the item correctly at the same rate as majority group students with comparable total scores. Large DIF values are an indication of additional difficulty, meaning that LEP students are having more difficulty on that item level than would be expected given their total score. In some cases, large DIF values should be ignored because they are based on very small numbers of LEP students. For example, levels 3 and 4 of item 4 appear to be much more difficult for LEP students than expected, but the results are based on only 16 and 10 LEP students, respectively.

Overall there was very little differential functioning for LEP students on the performance tasks. In some sense this is not surprising given that the language demands appeared to be uniform across all of the assessment tasks, with students being asked to explain their answers in every case. Thus whatever the effect of language on assessment of mathematics, it affected all items equally. Although written explanations are clearly central to the goal of having students communicate mathematically, it would be worthwhile to consider explicitly what weight the ability to explain should be given in the total mathematics score. Rather than using holistic scoring rubrics, analytic scoring methods, with separate points given for numeric answers and explanations, would make it possible to evaluate the effects of language more directly. Item 8 was the only item that showed substantial DIF across all four levels. It was a probability item involving the use of dice. It is possible that LEP students had not had experience with dice, and therefore might not know, for example, that only the top side of the die would count on any given throw or that different sides could turn up at different times.

Conclusions

The purpose of this study was to examine the effect of accommodations on the participation and performance levels of limited-English-proficient students in the Rhode Island Grade 4 Mathematics Performance Assessment. A pilot study was also conducted with a sample of 22 classrooms to provide preliminary evidence on the relative validity of both the performance assessment and the traditional Metropolitan Achievement Test for language-minority students compared to general education students.

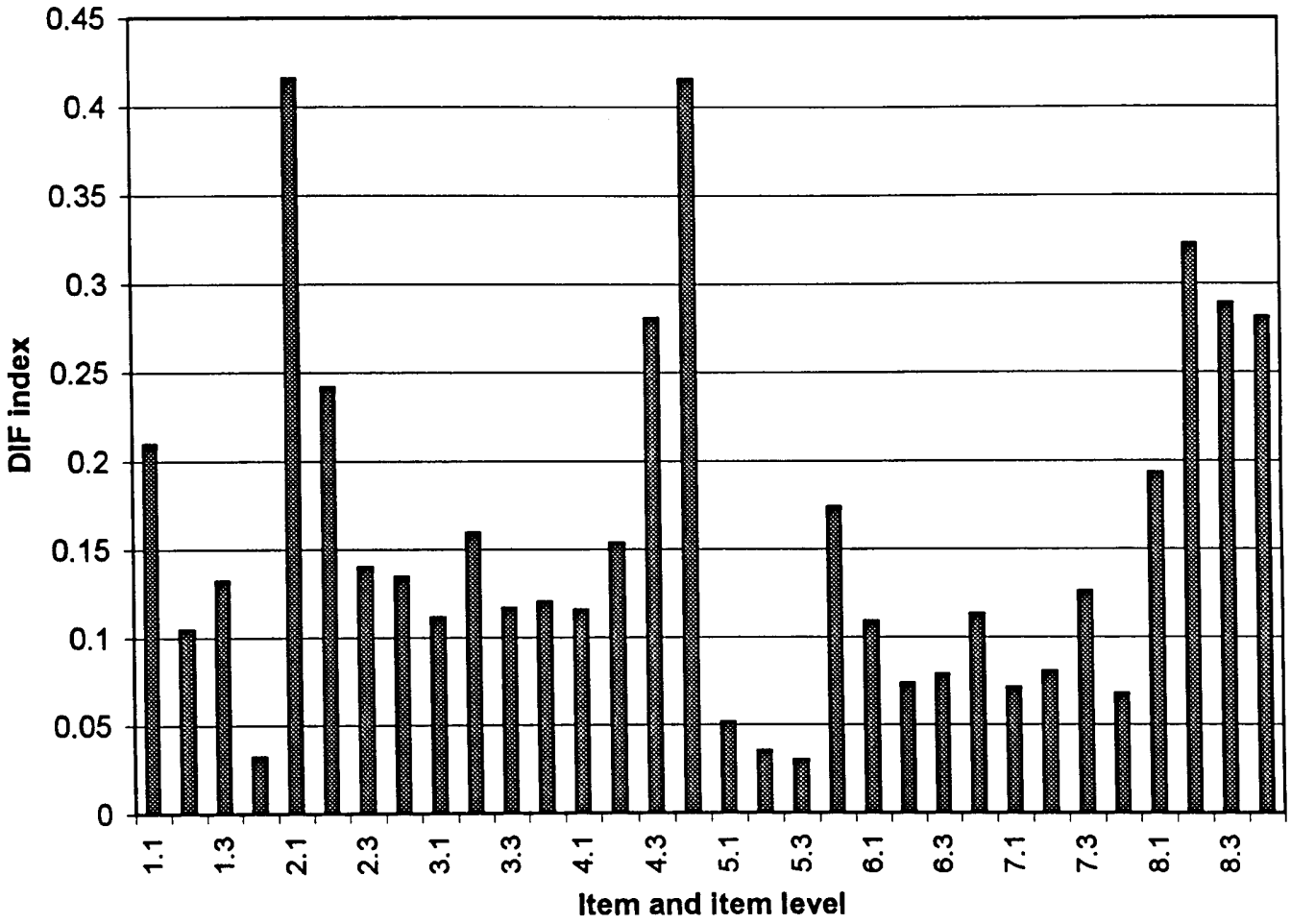

Figure 17. Amount of DIF for LEP students compared to state population on eight common items on the Grade 4 mathematics performance assessment.

In the statewide data, there was a clear increase in the numbers of LEP and special education students participating in the performance assessment compared to the number who took the MAT. This increase was most likely due to the availability of accommodations or to the accompanying directions that stressed the need for full inclusion. Accommodations consistently raised the relative position of LEP and special education students on the performance assessment compared to where they had been, relative to the general education mean, on the MAT. In the operational statewide assessment there was no way to evaluate the validity of achievement gains associated with the use of accommodations. For the most part, the level of gain appeared reasonable. However, there were examples of students who gained 1 or 1.5 standard deviations on the performance assessment compared to the MAT. Of particular concern was the finding that four schools had large numbers of LEP students who made these very large gains.

Very few LEP students received accommodations specific to their language needs. As has been found in previous studies, the vast majority of students receiving accommodations experience a change in the conditions of test administration: oral reading of the assessment, repeating directions, testing in a small group or ESL classroom, or receiving extended time. When the use of accommodations was examined by school, a troubling finding was that many schools accommodate "all or none" of their LEP and special education students. This suggests a greater need for training of school personnel so that they can make accommodation decisions more targeted to the needs of particular students.

In the pilot sample, teachers' mathematics grades and teachers' standards-based ratings in mathematics could be used as validity criteria to evaluate both the performance assessment and the MAT. The two tests were strongly correlated with each other and with the criterion variables suggesting that they do a good job of representing students' mathematics achievement. This promising validity picture was equally strong for language-minority students with advanced English proficiency. However, validity correlations were not as strong for students with limited English proficiency. In addition, it was noted that the designation of students as LEP was used inconsistently and represented a wide range of language levels.

For a first effort, the inclusion of LEP students in the Rhode Island Grade 4 Mathematics Performance Assessment appears to have been reasonably successful. Although there were clearly a small percentage of accommodated students who received inflated scores, the overall means were not implausibly distorted as Koretz observed in Kentucky. Better training is needed to make both better classification decisions (who is LEP?) and better accommodation decisions. It would be helpful to have more descriptive information to know, for example, why so few students were given Spanish language accommodations. Then validity data are needed on a wider sample to determine whether accommodated assessment results provide a more accurate picture of students' achievement. For policy purpose it would also be wise to keep track of non-native English speakers whose proficiency is no longer "limited." At present the performance of these students, after they graduate from ESL services, is lost in the data for general education students. Yet, it is the achievement of these students over time and across grades that ultimately reflects the success of second-language programs.

References

Abedi, J., Lord, C., & Plummer, J. R. (1997). *Final report of language background as a variable in NAEP mathematics performance* (CSE Tech. Rep. No. 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

August, D., & Hakuta, K. (Eds.). (1997). *Improving schooling for language-minority children: A research agenda*. Washington, DC: National Academy of Education.

Balow, I. H., Farr, R. C., & Hogan, T. P. (1993). *Metropolitan Achievement Test* (7th ed.). San Antonio, TX: The Psychological Corporation, Harcourt Brace Jovanovich.

Braun, H., Ragosta, M., & Kaplan, B. (1988). Predictive validity. In W. W Willingham, M. Ragosta, R. E. Bennett, H. Braun, D. A. Rock, & D. E. Powers (Eds.), *Testing handicapped people* (pp. 109-132). Boston, MA: Allyn and Bacon.

Butler, F. A., & Stevens, R. (1997). *Accommodation strategies for English language learners on large-scale assessments: Student characteristics and other characteristics* (CSE Tech. Rep. No. 448). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Council of Chief State School Officers and North Central Regional Educational Laboratory. (1996). *1996 state student assessment programs database*. Oak Brook, IL: North Central Regional Educational Laboratory.

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism, 19,* 197-205.

Fleischman, H. L., & Hopstock, P. J. (1993). *Descriptive study of services to limited English proficient students. Volume I: Summary of findings and conclusions*. Arlington, VA: Development Associates.

Koretz, D. (1997). *The assessment of students with disabilities in Kentucky* (CSE Tech. Rep. No. 431). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

LaCelle-Peterson, M., & Rivera, C. (1994). Is it real for all kids? A framework for equitable assessment policies for English language learners. *Harvard Educational Review, 64,* 55-75.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement*, *5*, 159-173.

McDonnell, L. M., McLaughlin, M. J., & Morison, P. (Eds.). (1997). *Educating one and all: Students with disabilities and standards-based reform*. Washington, DC: National Academy of Education.

National Academy of Education. (1996). *Quality and utility: The 1994 trial state assessment in reading*. Stanford, CA: National Academy of Education.

Olson, J. F., & Goldstein, A. A. (1997). *The inclusion of students with disabilities and limited English proficient students in large-scale assessments: A summary of recent progress*. Washington, DC: U.S. Department of Education.

Shepard, L. A. (1995). Assessment of language-minority students. In E. L. *Baker, Principal Investigator, Institutional grant proposal for OERI Center on Improving Student Assessment and Educational Accountability. Integrated assessment systems for policy and practice: Validity, fairness, credibility, and utility.* Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Shepard, L. A. (1996). *Research framework for investigating accommodations for language-minority students.* Presentation made at the annual CRESST Assessment Conference, UCLA, Los Angeles.

Stancavage, F., Allen, J., & Godlewski, C. (1996). Study of exclusion and assessability of students with limited English proficiency in the 1994 Trial State Assessment of the National Assessment of Educational Progress. In *Quality and utility: The 1994 trial state assessment in reading*. Stanford, CA: National Academy of Education.

Thissen, D. (1991). *MULTILOG User's guide: Multiple, categorical item analysis and test scoring using item response theory*. Chicago: Scientific Software.

Willingham, W. W., Ragosta, M., Bennett, R. E., Braun, H., Rock, D. A., & Powers, D. E., (Eds.). (1988). *Testing handicapped people*. Boston, MA: Allyn and Bacon.

Appendix: Methods of Analysis to Evaluate Differential Item Functioning

The eight composite performance tasks on the Rhode Island Mathematics Performance Assessment were analyzed using Multilog (Thissen, 1991), an IRT software program designed for test data with multiple response categories. Item responses were first analyzed using both LEP and non-LEP students in the same data set. Item parameters were estimated using a graded-response model with a random MML procedure. Multiple *b* parameters were found for each of the levels within a task. Item parameters were then fixed and individual theta scores were estimated for each student within the pooled data set. Then means and standard deviations were calculated separately for LEP and non-LEP students

In the second stage of analysis, item parameters and individual theta estimates were derived in separate analyses for the LEP and non-LEP data sets. Item parameters and theta estimates from the separate group analyses were then placed on the same scale using a linear transformation which adjusts for differences in means and standard deviations obtained in the pooled analyses, as follows:

$$\theta = c + d\theta^*$$

$$b = c = b^*(d)$$

$$a = a^* / d$$

The c and d constants were calculated as follows:

$$c = \theta - d\theta^*$$

$$d = S\theta / S\theta^*$$

Where θ, $S\theta$, *b*, *a* equal the group mean, standard deviation, and *a* and *b* parameters of either the LEP or the non-LEP students in the joint data set, and θ^* $S\theta^*$ b^* a^* equal the group mean, standard deviation, and *a* and *b* parameters of the corresponding group in the separate data sets.

The following probability index suggested by Linn, Levine, Hastings, and Wardrop (1981) was used to find the area between the item characteristic curves for LEP and non-LEP students.

$$A_{2i} = \sum \{[P_{i1}(\theta_k) - P_{i2}(\theta_k)]^2 \Delta\theta\}^{1/2}$$

Where P_i are the probability levels for the respective groups at each .5 theta increment.