Part Two

Constructive Uses of Tests

Chapter 3

Early Reading Assessment

Barbara R. Foorman,
Jack M. Fletcher, and
David J. Francis

Early Assessment

When confronted with the large reported numbers of people who are reading-disabled (up to 17.5 percent nationally)[1] and poor readers (more than 64 percent of African-American and 60 percent of Hispanic children, according to the fourth-grade National Assessment of Educational Progress [NAEP]), policy makers want to know why these numbers are so large and whether reading problems can be prevented. These concerns prompted the U.S. Department of Education and the U.S. Department of Health and Human Services to establish a committee through the National Research Council to investigate the prevention of reading difficulties. The resulting report, *Preventing Reading Difficulties in Young Children,*[2] focuses on the conditions under

Supported by grants from the National Institute of Child Health and Human Development, R01 HD30995, "Early Interventions in Children with Reading Problems" and P01 21889, "Psycholinguistic and Biological Mechanisms in Dyslexia."

which reading success is likely to emerge. Success comes when teachers teach reading in a comprehensive way that emphasizes the importance of letter-sound relations and reading for meaning and that provides opportunities to practice. Those most at risk for reading difficulties are, as a group, children from low-income families. They live in poor neighborhoods, attend schools with low achievement, have limited English proficiency, and speak a dialect of English substantially different from the one spoken at school.[3] Researchers also have found that in addition to such family background, there are individual risk factors of reading difficulty: limited experience at home with reading and physical, language, or cognitive weaknesses involving "cognitive-linguistic processing, especially phonological awareness, confrontation naming, sentence/story recall, and general language ability."[4]

Given that the National Research Council report represents the consensus of empirical researchers on predicting success and failure in reading, how should the classroom teacher apply this knowledge? How does the kindergarten, first-grade, or second-grade teacher know which students are headed for reading success and which for reading failure? One answer is provided through early assessment of reading growth and outcomes.

In this chapter we divide the topic of early reading assessment into the following sections: (1) the importance of assessing early reading skills; (2) impediments to early reading assessment; (3) "formal" and "authentic" early reading assessments; and (4) the example of the Texas Primary Reading Inventory.

The Importance of Assessing Early Reading Skills

When children exhibit reading problems at an early age, these problems typically persist. There is little evidence that they catch up in reading skills, in spite of the widespread belief among educators in developmental delay—the late bloomer phenomenon. In one study from our group, 74 percent of children who were reading disabled in the third grade remained reading disabled in the ninth grade.[5] In fact, the presence of risk characteristics is apparent in kindergarten and grade one. Juel found that 88 percent of students who were poor readers in grade one were also

poor readers in grade four; 87 percent of students who were good readers in grade one were also good readers in grade four.[6] (But see Phillips and colleagues for a different perspective.[7]) In short, first grade matters in determining a child's status as a reader. Torgesen[8] found similar stability in reading status from grade one through grade five, but this status was predictable based on kindergarten performance—confirming that kindergarten matters as well.

The good news is that recent research indicates that early intervention is effective. Torgesen and colleagues identified children in kindergarten who had poor phonological awareness, that is, they had difficulty blending and segmenting sounds in speech.[9] By second grade, one-on-one tutoring brought 75 percent of the children to grade-level reading. Vellutino and colleagues identified middle-class children with very low word recognition skills at the beginning of grade one.[10] After one semester of one-on-one tutoring, 70 percent were reading at grade level. After two semesters, more than 90 percent were at grade level.

Tunmer and colleagues show the benefits of adding explicit, systematic alphabetic instruction to Reading Recovery tutorials.[11] Foorman and colleagues found that classroom-level explicit instruction in phonological awareness and the alphabetic principle as part of a balanced approach to reading brought students in grades one and two in eight Title 1 schools to national averages.[12] Less explicit, inductive approaches were unable to show such gains.

In sum, it is important to assess reading skills early for three reasons. First, reading status is a stable characteristic from as early as first grade and becomes intractable after third grade. Second, the presence of risk characteristics is apparent in kindergarten and first grade. Third, early intervention in first and second grade is effective.

Impediments to Early Reading Assessment

The major impediment to assessing reading skill development early in school is the "wait and see" attitude apparent in many

areas of early education. Consequently, children have to accumulate sufficient failure on standardized achievement tests administered in second or third grade before they are eligible for special education testing. Identification of "learning disabilities"—the label under which children with reading disabilities are typically served—is largely based on a significant discrepancy between IQ and reading achievement. In the 1950s, Bond and Tinker argued in favor of an IQ-discrepancy definition, noting that a child with reading disability "is a child who is not reading as well as could be expected for one of his intellectual or verbal maturity."[13] They further indicated that children with IQ scores below 95 should not be considered disabled in reading because "they are reading about as well as can be expected in view of their limited intellectual ability."[14] Much of the broad acceptance of IQ discrepancy models was fueled by the Isle of Wight studies by Rutter and Yule.[15] These studies presumably showed that reading skills had a bimodal distribution, with a longer tail representing children who were generally behind in reading relative to age but not IQ (low achievers). These children were contrasted to children with "specific" forms of reading disability reflected by poor reading in relation to expectations based on IQ (discrepant). These two groups of children were further shown to differ in gender, prognosis, reading and spelling characteristics, and language skills.

None of these findings has held up in research in the 1980s and 1990s.[16] The viability of IQ-discrepancy models is widely questioned. This has major implications for existing policy because the implementation of special education standards for children with reading disabilities still makes use of an IQ-discrepancy model. Interestingly, these concerns about IQ-discrepancy have come about despite major improvements in how IQ-discrepancy is modeled. For example, the types of definitions used in the 1950s, which typically involved a grade-below definition, have been widely questioned because of problems with their psychometric properties.[17] There is a significant literature on measuring IQ-discrepancy. The bulk of the evidence found in the literature clearly indicates that regression-based models that adjust for the correlation of IQ and achievement are the most

appropriate.[18] However, much of the research simply does not show major variations in the phenotypic characteristics of reading disability according to any definition.

In the last decade there has been much more attention paid to the assessment of domain-specific skills that are related to reading disability and a general deemphasis on the role of IQ tests.[19] This has happened largely because of a shift from organismic neuropsychological models to cognitive models of reading disability. There are multiple cognitive models of reading (for example, dual route, connectionist, and interactive). These models have been applied to reading disability with varying degrees of success.[20] In addition, research in the 1990s reflects the emphasis on the need for larger, well-defined samples, stronger hypothesis formulation, and the importance of a multivariate rather than univariate approach to research design. The following characteristics of the disabled reader have emerged from this recent research:[21]

1. Reading problems in most children occur at the level of the single word and involve word recognition skills. The best predictor of poor reading comprehension skills is deficient word recognition ability. Text reading problems account for far fewer cases of reading disability than problems with the development of word recognition skills.[22]
2. Word recognition problems are primarily associated with difficulties in segmenting words and syllables into phonemes. Deficits in phonological awareness characterize most poor readers, whether they are children, adolescents, or adults (at all levels of intelligence) or from economically disadvantaged or non-English speaking backgrounds. Individuals with reading disability have difficulty mapping speech into print.[23]
3. Reading ability lies along a natural, unbroken continuum. There is no natural demarcation on this continuum that separates good and poor readers, and a major research topic is the severity of impairment of reading skills that constitutes a disability. Most of the more current research in the

area of reading disabilities operates, at least implicitly, from a dimensional rather than a categorical model.[24]

4. Reading difficulties occur with equal frequency in boys and girls. The puzzle is that schools identify four times more boys than girls.[25] Vernon said that this was because of boys' behavioral characteristics.[26] He noted that reports of differences between the sexes were most likely related to the use of clinic samples, and, like Shaywitz and colleagues,[27] he further noted that boys were "most resistant to school teaching and discipline"—resulting in referral to a clinic.
5. As stated early in this chapter, children's early reading problems typically persist throughout their schooling. There is little evidence that they catch up in reading skills.
6. Reading disability is best identified through domain-specific assessments of reading and reading-related skills.[28] IQ tests are not necessary for the identification of reading disability. Models for identification based on IQ discrepancy lack validity. Stated another way, there is little evidence for differences between IQ-discrepant and low-achieving children on multiple dimensions, including the cognitive characteristics of reading disability, response to intervention, and long-term outcomes.[29]
7. There are multiple distal causes of reading disability, including (a) neurological; (b) familial; (c) economic (low income) and linguistic (low English proficiency and dialect differences); and (d) instructional:

   (a) Because of improvements in definition and measurement, major advances have been made in understanding the neurobiological correlates of reading disability. In particular, researchers have studied brain metabolism using positron emission tomography and functional magnetic resonance imaging. Such research shows that when adults with reading disability complete word-recognition tasks that separate the phonological, orthographic, and semantic components of word recognition, several of the neural networks that

they use are different from those used by good readers.[30] Previous studies had reported problems with brain structure, but these problems are subtle, and differences are less robust than differences apparent in studies of brain function. Although researchers in the 1950s were interested in brain damage as a cause of phonological reading disability, it is clearly not a major cause. In fact, brain injury often results in preservation of word recognition skills and is more likely to lead to impairment at the level of text processes.[31]

(b) It has long been known that reading problems run in families. Recent studies have shown specific genetic loadings at chromosome 6 and 15.[32] However, several combinations of genes appear to be involved and the penetrance is low. Environmental factors clearly have significant influences on reading outcomes, and heritability accounts for about 50 percent of the variability.

(c) In addition to research on neurobiological factors, recent research also clearly establishes the importance of environmental factors for the development of reading disability. This is apparent in the large number of minority children with poor reading achievement on the National Assessment of Educational Progress (NAEP). The child's early literacy environment has a significant influence on reading outcomes.[33]

(d) The influence of instructional factors is underestimated. Recent studies have shown that intervention, particularly if it is early, can succeed in improving the word recognition skills of children with reading disability.[34]

8. Studies show that interventions that focus specifically on word recognition skills appear to help children overcome the most common forms of reading disability. The best available data suggests that this instruction needs to be explicit, must emphasize the alphabetic principle, and requires some intensity. Success has been reported both in

> classroom level studies and in tutorial and pullout models.[35] The interventions that are successful provide an explicit focus on alphabetic decoding and word recognition skills, but are also characterized by an emphasis on reading connected text, writing, and reading then discussing intellectually challenging literature. This reflects a broad view of the reading process and the importance of applications of skills, particularly for children with reading disability. It is clear that many children with reading disability read and write less than other children, particularly if they are in traditional school-based remedial programs. Outcomes in any reading program are tied to the amount of practice, so it is important to get children to read and write and apply the skills they learned in the intervention program.

This summary makes clear that oral language skills and reading skills are related. The relationship does not apply solely to children who have speech and language disorders, although this is a population that is at substantial risk for the development of reading problems. Recent epidemiological studies suggest that approximately 50 percent of children with a history of an oral language disorder will develop a reading disability.[36]

Language factors do appear to account both for success and failure in the acquisition of reading skills. This point, argued most persistently by Vellutino in the 1970s,[37] is clearly the predominant view today among reading researchers. Researchers understand that reading is an unnatural outgrowth of language. Reading and writing are scaffolded onto oral language. Children do not acquire reading skills through exposure to literature, and reading does not develop naturally as does oral language. Reading must be taught, and the component of reading that requires the most explicit instruction is the relationship between print and speech at the level of the single word. The pivotal role of phonological awareness skills in the acquisition of reading ability is now well established and represents the most robust proximal cause of reading failure.[38]

Much of this summary of research on reading disability is based on children and adults identified as "reading disabled." Yet there is little evidence that the causes of reading problems in chil-

dren who come from environments with limited literacy exposure are different from the causes of reading problems in children who come from non-English speaking backgrounds. In recent research on intervention with samples in which children from linguistically and culturally diverse backgrounds predominated, the evidence for word-level deficiencies in the early grades and the mediating effects of improvements in phonological processing skills is impressive.[39] This is not to say that word-level skills are all that teachers need to address or that addressing these skills will eliminate reading failure. Rather, the point is that the development of word-level skills is necessary—but not sufficient—for preventing reading failure. Instructional programs at all levels must integrate alphabetic instruction with opportunities to read connected text and an emphasis on meaning. Nonetheless, this summary makes it clear that we know a great deal about the characteristics of the disabled reader. We know the importance of prevention and early intervention. The question becomes how to identify children at risk for reading disabilities so that instructional support may be provided early on.

Formal and Authentic Reading Assessments

In the past, testing of early literacy was remarkably unconnected to teaching practices, although some claimed that testing harmed teaching and learning.[40] The source of this prevalent disconnection was the schism in the discipline of psychology between behaviorists and rationalists. As a result, on the one hand, a professional class of testers administered formal tests of skills as part of the accountability system, whereas on the other hand, constructivist teachers spent time filling out informal inventories. Those using either of these approaches typically missed the central purpose of assessment: setting individual learning objectives on the basis of systematically gathered information. In the following section, we discuss examples of "formal" and "authentic" assessment of early reading and then examples of attempts to merge the two.

Formal reading tests

We use the term "formal" to designate tests that are part of the accountability system (and, therefore, are high stakes). Formal

tests can be norm-referenced or criterion-referenced. Norm-referenced tests are standardized on a clearly defined group, termed the norm group, and scaled so that each individual score reflects a rank within the norm group.[41] Criterion-referenced tests rate students against the content being assessed.[42]

Formal reading assessments are generally well-known, so are not be described in detail here. The Stanford Achievement Test-10 is an example of a well-known group-administered, norm-referenced achievement test. The Woodcock-Johnson[43] is a well-known individually administered achievement test. In both cases, students' word identification and passage comprehension scores are reported in relation to students of the same age or grade. In contrast, Texas uses the Texas Assessment of Knowledge and Skills (TAKS), a group-administered criterion-referenced test, as its accountability device. Scores on the TAKS reflect the percentage of items passed.

All of these formal tests use a multiple-choice format. In an attempt to go beyond the declarative knowledge tapped by multiple-choice formats, performance-based assessments, in which the student constructs an original response (that is, displays procedural knowledge) and the examiner observes the process of construction, have become popular alternatives. Assessment of writing fits well into this format. Reading comprehension also fits well if longer passages are used and students are asked to write responses to questions rather than to select the best alternative. The NAEP and the New Standards Project[44] are examples of national performance-based assessments.

Performance assessments that measure children's mastery of specific curriculum objectives through formal and informal approaches are not discussed in this chapter. The many curriculum-based assessments found in basal reading programs used to monitor student progress and make placement decisions are examples of this kind of performance assessment. These assessments rarely present evidence of reliability or validity and generally do not measure transfer of knowledge independently of the specific curriculum. Interestingly, hybrid approaches do exist that measure students' mastery of a curriculum and transfer of knowledge. One notable example, that of Fuchs and Fuchs, also has

excellent reliability and validity, with a strong empirical basis of support.[45]

For the present discussion, we see no inherent reason why the value of the procedural knowledge tested by performance assessment would negate the value of declarative knowledge. For example, it is important to assess vocabulary knowledge prior to instruction in reading. Yet, in order to construct an original response, the student needs to write definitions or give them orally. Written and oral responses are complex tasks all by themselves because they require more than just vocabulary knowledge. Moreover, scoring of written and oral responses is time-consuming and often unreliable. In fact, the judicious use of multiple-choice formats to assess declarative knowledge may be the most valid, reliable, and useful way to proceed. Thus, performance assessment should be used as an addition to rather than a replacement for more traditional formats.[46] Similarly, performance-based assessments and authentic assessments can be complementary. The research literature provides several examples of combined assessments of reading and literacy skills.[47] Linn and colleagues remind us, in their writings on complex, performance-based assessment, that:

> Serious validation of alternative assessments needs to include evidence regarding the intended and unintended consequences, the degree to which performance on specific assessment tasks transfers, and the fairness of the assessments. Evidence is also needed regarding the cognitive complexity of the processes students employ in solving assessment problems and the meaningfulness of the problems for students and teachers. In addition, a basis for judging both the content quality and the comprehensiveness of the content coverage needs to be provided. Finally, the cost of the assessment must be justified.[48]

Authentic assessment. Performance assessments are "authentic" to the degree that they "reflect the actual learning and instructional activities of the classroom and out-of-school worlds."[49] The term "authentic assessment" covers a wide territory with great variability in how formal the test procedures are. Some authentic assessment is rather formal, and some is quite informal. At the least formal end of the continuum are portfolios of student

work. The use of portfolios that has proved most enduring is as a component of local assessment systems agreed upon by parents, teachers, administrators, and school board members. Six examples of authentic assessment systems, some of which are implemented on a large-scale basis, are described here.

1. *Observation Survey*[50] This diagnostic battery of tests, developed in New Zealand and used in the Reading Recovery program in that country and in the United States, is the blueprint for many current authentic literacy assessments. Part I of the battery uses a technique called "running records." A teacher listens to a child read and takes notes ("running records") of oral reading errors and self-corrections. Part II of the battery includes letter identification and concepts about print, word reading, writing, and dictation. No inter-rater reliability is reported for the running records in Part I; however, excellent reliability is reported for the tests in Part II in multiple studies with different sample sizes and compositions: test-retest, .73–.98; internal consistency, .84–.97. Concurrent validity is reported as correlations of letter identification with word reading (r = .96), concepts about print with word reading (r = .79), and word reading with the Schnoell Reading-1 Test (r = .90). Evidence for predictive validity can be found in a study reporting that Reading Recovery brought 35 percent of children served through the one-on-one tutorial to the classroom average.

   Marie Clay's work is commendable for its attention to issues of validity and reliability, authenticity, and professional development.[51] Becoming a Reading Recovery tutor requires one year of intensive clinical work. Tutors have to learn to observe children, code running records, administer the diagnostic tests, and apply results to instructional planning. The expense of year-long teacher training and one-on-one tutoring has resulted in many adaptations of the model under other names. These adaptations typically include the same 30-minute lesson cycle for tutorials: rereading of a previously read book; independent reading of

a newly introduced book with a running record taken; letter identification exercises, if necessary; writing and reading of sentence strips; cutting up and reassembling words on sentence strips; and introduction of a new book with scaffolded reading.

Program developers of Reading Recovery report positive long-term gains on concepts of print and dictation.[52] However, external evaluations do not report strong transfer to other reading measures.[53] Moreover, Iversen and Tunmer found that children were "recovered" at a faster rate if the lesson cycle included systematic instruction in letter-sound patterns.[54] Finally, Center and colleagues[55] and Tunmer and colleagues[56] found that children with poor metalinguistic knowledge were less likely to be successful in Reading Recovery. That may explain why approximately 27 percent of children served are dismissed from the program without being recovered.[57]

2. South Brunswick, New Jersey, Schools' *Early Literacy Portfolio*[58] This suburban district of seven elementary schools developed a portfolio system for children in kindergarten and first and second grades that consisted of these components:
   - Writing samples
   - Story retelling records
   - Oral reading records
   - An invented spelling activity
   - Sight word inventories
   - Interviews with parents and students
   - Self-portraits

   Teachers collected documentation for each of these components at the middle and end of each year. They rated the quality of the documentation according to a 6-point scale referenced to expected literacy performance at each phase of development. These ratings were used to monitor student

progress and to meet local and state evaluation requirements. Each year teachers met across schools to blindly rate each other's portfolios in order to check inter-rater reliability of scoring. Reliability has been high (.90s).

Salinger and Chittenden[59] interviewed teachers about their use of the portfolio system. The profile of a student's strengths and weaknesses across the seven components mattered more to teachers than the developmental scale. This profile helped teachers plan for instruction and for meetings with parents. Teachers' biggest complaints concerned management and time.

Compiling multiple documents for each child at the middle and end of the school year requires organization and time. Moreover, some experienced teachers admitted that the portfolio was redundant with what they already knew about their students through instruction. But for new and developing teachers, portfolios provide a mechanism whereby they spend individual time with each student and reflect on the impact of their instruction on literacy development.

3. *The Primary Language Record,* developed in London and used at P.S. 261 in New York City,[60] consists of writing samples, running records, interviews with parents and students, and classroom observations. In California the PLR has been adapted by classroom teachers and renamed the California Learning Record(CLR).[61] Five developmental levels of reading proficiency have been defined for kindergarten through third grade, fourth through eighth grade, and high school.

   The K-3 scale is described in the 1996 moderation report:[62]

   a. Beginning reader: Uses just a few successful strategies for tackling print independently. Relies on having another person to read the text aloud. May still be unaware that text carries meaning.

b. Not-yet-fluent reader: Tackles known and predictable text with growing confidence but still needs support with new and unfamiliar ones. Has a growing ability to predict meanings and developing strategies to check predictions against other cues, such as the illustrations and the print itself.

c. Moderately fluent reader (partially proficient): Well launched on reading but still needs to return to a familiar range of reader text. At the same time, is beginning to explore new kinds of texts independently and is beginning to read silently.

d. Fluent reader (proficient): A capable reader who now approaches familiar texts with confidence but still needs support with unfamiliar materials. Is beginning to draw inferences from books and stories. Reads independently. Chooses to read silently.

e. Exceptionally fluent reader (advanced): An avid and independent reader who is making choices from a wider range of material. Able to appreciate nuances and subtlety in text.

Inter-rater reliability for placement of students into these levels of proficiency was 85 percent within a school site and 70 percent to 80 percent across sites within a region. Teacher reports are used as evidence of the impact of the CLR upon classroom instruction and are collected annually as part of the moderation process.

4. *The Primary Assessment of Language Arts and Mathematics (PALM)* was designed in Austin, Texas, by researchers and teachers to include three components: (a) curriculum-embedded assessments, (b) taking-a-closer-look assessments, and (c) on-demand assessments.[63] The curriculum-embedded assessments consist of the ongoing gathering of evidence to document progress in the curriculum. This evidence consists of work samples, classroom observations, and anecdotal records. The taking-a-closer-look assessments include

informal reading inventories, running records or miscue analyses of oral reading, and think-aloud and reflective problem-solving strategies—all tools that teachers might use to gain further information about individual students. On-demand assessments for the PALM include: a personal journal; a response journal for both a book read aloud by the teacher and a free-choice book read independently; an adaptation of the K-W-L (know, want to learn, learned) model for expository text;[64] an oral reading of a familiar and an unfamiliar passage scored for accuracy, rate, and self-correction; and an inventory of reading attitudes and habits and of self-concept.

Hoffman and colleagues found that teachers could implement the PALM and use the results to plan instruction.[65] However, no evidence of reliability is provided. Evidence of concurrent validity was established with the Iowa Test of Basic Skills (ITBS). Hoffman and colleagues also found that the PALM accounted for 86 percent of the variance in the ITBS reading score.[66] However, as Pearson points out, such overlap with the ITBS may not be a blessing.[67] Given the time-consuming nature of the PALM, one might argue that the ITBS is a cost-effective substitute. The issue is an empirical one: Which assessment approach has more utility—the PALM or the ITBS—in assuring that students become successful readers one year and two years later? This is an issue both of predictive validity and diagnostic utility. In other words, which approach reliably identifies students in need of additional assistance if they are to become successful readers and which approach provides teachers with information about the nature of the assistance needed? The PALM clearly provides information directly relevant to the content of instructional assistance. However, a longitudinal study of student growth and outcomes is needed to address whether the PALM has predictive validity.

5. *The Work Sampling System*[68] Meisels's Work Sampling System involves elementary school teachers' documenta-

tion and evaluation of ongoing student work with the goal of improving instructional practices and student learning. Three forms of documentation are used: checklists, portfolios, and summary reports. The checklists consist of performance indicators for seven major curriculum areas drawn from national and state curriculum standards. For example, "Understands and interprets a story or other text" is one indicator from the first-grade checklist. Teachers check the three-level mastery scale—Not Yet, In Process, and Proficient—in fall, winter, and spring assessments to trace student performance. Detailed developmental guidelines accompany each checklist area in order to promote consistency of interpretation and evaluation across teachers, students, and schools.

In Meisels's system, portfolios consist of two types of student work: core items and individualized items. Core items represent performance in five domains—language and literacy, mathematical thinking, scientific thinking, social studies, and the arts. Individualized items reflect a child's goals, interests, and abilities in various curricular areas, such as first attempts at acrylic painting or writing a story. The inclusion of core items provides for structured sampling of performance across students. Individualized items provide the opportunity to represent student strengths and to enable students to take an active role in evaluating their own work.

Summary Reports transform information from teacher observations, checklists, and portfolios into evaluation of student performance across the curriculum. Teachers complete these reports three times a year, writing an evaluation in narrative form and completing a rating scale for each of the five domains. The ratings are: (1) not yet accomplished, (2) accomplished, or (3) highly accomplished. A total summary score is created by summing ratings across domains and across the three subscales—observations, checklists, and portfolios.

Meisels and colleagues examined the Work Sampling System's reliability and validity with 100 kindergartners.[69] Results showed that the checklist and summary report (including portfolio ratings) had high internal and moderately high inter-rater reliability. Also, the Work Sampling System accurately predicted performance on the norm-referenced battery of individually administered achievement tests, controlling for sex, age, and initial ability.

6. *Phonological Awareness and Literacy Screening (PALS)* PALS was developed at the University of Virginia with funds from the state in order to develop a tool that teachers could use to identify kindergarten and first-grade students who might benefit from additional instruction. There are two parts to PALS—phonological awareness (PALS I) and literacy screening (PALS II). PALS I assesses ability to identify rhyme units and to isolate beginning sounds, in an individual or small group format. PALS II assesses (1) alphabet knowledge, (2) knowledge of letter sounds, (3) concept of word, (4) sense of story, and, in first grade, (5) word recognition. Letter knowledge is assessed through recognition of upper- and lowercase letters and production of a subset of letters. Knowledge of letter sounds is assessed through (a) production of letter sounds in isolation, (b) ability to categorize beginning sounds, and (c) ability to use knowledge of letter sounds to attempt to spell. Concept of word is measured by ability to track words in familiar text as well as ability to use context to identify individual words within a line of text. Sense of story is measured through story retelling. Word recognition in first grade is assessed with graded word lists.

   University of Virginia researchers received PALS scores from 52,094 kindergarten and first-grade children in the 1997–1998 and 1998–1999 school years, with more than 90 percent of school divisions in Virginia returning data. Item reliabilities were determined for grade, gender, socioeconomic status, and geographical region, yielding Cronbach's Alphas

ranging from .83 to .89. Ethnicity was available in the 1998–1999 administration. Inter-rater reliability of .99 for each subtest was obtained when teams of two adults (not the actual teachers) administered the PALS screening to the same children in six schools across three regions in Virginia in the fall of 1999. Construct validity was addressed through factor analysis of the 1997 PALS data. Both kindergarten and first-grade data were best represented by a single-factor solution that accounted for 64 percent to 74 percent of the total variance in the children's scores on all tasks in both the phonological awareness and literacy screening components. The subtasks contributing the most to the one-factor solution were rhyme, beginning sounds, lowercase alphabet recognition, letter sounds, and spelling. These were retained in the current version of PALS. Of these five subtasks, lowercase alphabet recognition, letter sounds, and spelling contributed the most to the unitary factor. Concurrent validity was established with medium to high correlations (.67 to .81) with Stanford-9 subtests of sounds and letters, word reading, and sentence reading administered to 127 first-graders in the fall of 1997.

These examples show that a number of authentic assessment systems are available. Mostly, they are local efforts to engage teachers in collecting evidence upon which to base individual curricular decisions. Three of the six systems presented are large-scale applications—the CLR, the Work Sampling System, and PALS. The latter two systems have been the most responsive to psychometric concerns regarding validity and reliability. Pearson regards validity as the ultimate criterion for judging the worth of a test and, therefore, for judging the worth of an assessment system.[70] He lists the following as the important questions to ask in determining the validity of a test:

1. Does it measure the intended trait? (construct validity)
2. Is it consistent with the curriculum? (content validity)

3. Does it behave like other measures of this domain? (concurrent validity)
4. Does it result in appropriate decisions for users? Do they get what they need? (consequential validity)
5. How much effort is required to obtain the information? (feasibility)
6. How do users judge the quality and appropriateness of the information they receive? (utility)[71]

In judging the validity of an assessment system, Pearson raises these three questions:

1. Are all of the important dimensions of the domain assessed? This question speaks to the issue of domain or content validity. The items being from the appropriate domain is not enough to establish their system validity. For the system to be valid, the entire domain must be adequately represented.
2. Are the clients of the system getting the information they need in order to answer the questions they want answered? This question speaks to the criterion of utility and emphasizes the "tailoring" of the information to the audience who will use it.
3. Are clients making the right decisions? This question addresses issues of consequential validity. It must be answered by examining the impact of such assessments on the lives of individuals and groups who are affected by the results of the assessments. The ultimate test is whether appropriate placements and instructional decisions are made. Particularly important to examine are egregious misapplications of the system; other things being equal, we want assessments that do no harm.

These are important and clearly articulated aspects of validity, but Pearson fails to list one kind of validity important to early reading assessment—predictive validity. Does the test or the sys-

tem predict future reading performance? To address predictive validity, longitudinal studies of individual growth and outcomes in reading are required. Currently, the only early reading assessment with evidence of predictive validity is the Texas Primary Reading Inventory (TPRI).[73] However, let us address the notion of reliability, for it is conspicuously missing from Pearson's discussion.

"Reliability" measures the consistency or reproducibility of test scores. In practical terms, reliability is the extent to which a student's score remains constant when the same test is given under a variety of conditions. The reliability of an instrument is important in school settings because educators and parents want to make sure that a student's score is representative of the student's ability and not a reflection of random error. "Internal consistency" provides an estimate of the error in using the subset of items on the test instead of using all possible items from the domain of items. "Alternate forms reliability" estimates the error in using two forms for measuring the same trait. "Test-retest reliability" is an estimate of the error associated with testing over time. "Inter-rater reliability" reflects the consistency with which different raters score a student.

Recently, some researchers have reduced the importance of reliability relative to validity. For example, Tierney offers thirteen principles of assessment,[74] one of which is: "Some things that can be assessed reliably across raters are not worth assessing; some things that are worth assessing may be difficult to assess reliably except by the same rater" (384). Difficult as it may be to achieve high inter-rater reliability, the concept of consistency and reproducibility is essential if a test is to be considered valid. If classroom teachers are inconsistent in setting learning objectives based on assessment results, then the validity of the assessment instrument can be questioned. In short, validity can be no stronger than reliability. A test can be reliable, but not valid, as the first part of Tierney's statement says. However, tests can never be valid and unreliable. Hence, we urge educators not to abandon the notion of reliability just because it may be hard to achieve. Rather, we urge educators to gather evidence of reliability so that we can fully address issues of validity.

Texas Primary Reading Inventory

All school districts in Texas are required to administer an early reading diagnostic instrument for students in kindergarten, grade one, and grade two according to Texas Education Code 28.006. This requirement developed out of the 75th Texas Legislature with the passage of House Bill 107 in May 1997. Texas Education Code 28.006 is explicitly *not* part of the accountability or teacher appraisal or incentive system in Texas. Assessment results are to be reported to parents, superintendents, school boards, and the Commissioner. The state does not mandate what assessment is used, but does provide support for assessments on a list from the Texas Education Agency that includes instruments that can be individually administered by a teacher and that have evidence of reliability and validity.

In order to facilitate this mandated diagnosis of early reading skills and comprehension development, the Texas Education Agency (TEA) contracted with the Center for Academic and Reading Skills to revise an early diagnostic reading instrument developed by the TEA known as the Texas Primary Reading Inventory. There are more than 1,000 school districts in Texas, with almost one million children in kindergarten, first grade, and second grade, taught by more than 45,000 teachers. During 1998–1999—the first year of implementation of Education Code 28.006—approximately 80 percent of school districts adopted the TPRI. During 1999-2000, 85 percent of school districts adopted the TPRI and many piloted a Spanish reconstruction called the Tejas LEE. During 2000-2001, over 90 percent of school districts adopted the TPRI and the Tejas LEE, and in 2001–2002 the percentage rose to 95 percent. In the 2004–2005 edition, a third-grade screen and inventory and a progress-monitoring booklet were added to meet the requirements of the Reading First component of the ESEA's No Child Left Behind legislation. The TPRI was developed as a large-scale example of an early reading instrument that attempts to bring psychometric rigor to informal assessment, as PALS in Virginia attempts to do. In this section we will describe in detail the development and implementation of the TPRI.

Development of the TPRI

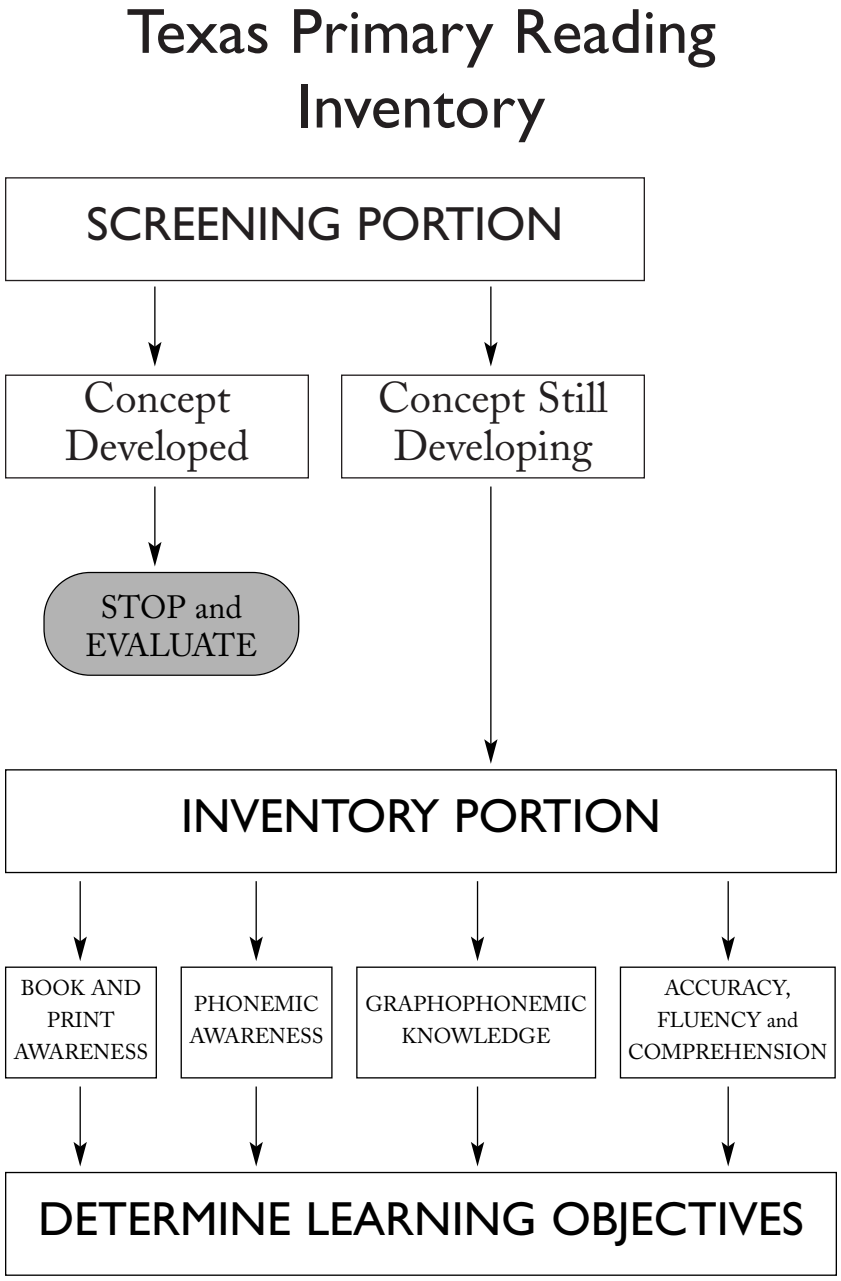
In revising the TPRI, we (a) added a screening component to identify those students who had high probabilities of success at the end of grades one and two, and (b) modified the inventory portion to be aligned with the new state curriculum standards and to be more easily scored by teachers. The screen consists of those measures most predictive of reading success in our longitudinal sample of more than 900 children in kindergarten through grade two and parallels closely the work of Torgesen and Vellutino and colleagues.[75] These measures are: phonological awareness and its theoretically related construct of letter-sound knowledge in kindergarten and the beginning of grade one; and word reading at the beginning and end of grade one and beginning of grade two. For children still developing these screening concepts, the inventory is administered to set learning objectives. These administration procedures are described in Figure 3.1.

The components of the TPRI are defined in the teacher guides for kindergarten, first, and second grades as follows:

- Book and print awareness – knowledge of the function of print and of the characteristics of books and other print materials
- Phonemic awareness – the ability to detect and identify individual sounds within spoken words
- Graphophonemic knowledge – the recognition of the letters of the alphabet and the understanding of sound-symbol relationships
- Reading accuracy and fluency – the ability to read grade-appropriate text accurately and fluently
- Reading comprehension – the understanding of what has been read

All tasks in the TPRI consist of five questions, with concept development indicated by four out of five correct. The book and print awareness task, inspired by Clay's Concepts of Print,[76] asks the teacher to select a short storybook and to ask the child to

FIGURE 3.1 Flowchart of components of the TPRI



point to the place where the teacher starts reading, to point to the start and end of a sentence, and to point to a word, a letter, and a capital letter. This task is included in the kindergarten inventory as a warm-up activity. It is not scored because of lack of inter-rater reliability in the validation study.[77]

The phonemic awareness tasks on the kindergarten inventory are (a) rhyming, and (b) blending word parts. In the rhyming task, the teacher gives the child three rhyming words (for example, hill, fill, dill) and the child is to produce a word, real or made up, that rhymes with these three. In the second task, the teacher pronounces a single-syllable word broken into the initial sound (onset) and the final pattern (the rhyme), such as /h/-/ouse/ or /ch/-/in/. The child's job is to put the word-parts back together. These two tasks also appear on the first-grade inventory. Additional phonemic awareness tasks on the first-grade inventory are blending phonemes in spoken words (/s/-/u/-/n/→"sun") and detecting initial sounds (say "sit" without the /s/→"it") and final sounds (say "beef" without the /f/→"bee").

The graphophonemic knowledge tasks vary in format from kindergarten through grade two. In kindergarten there are two tasks—letter-name identification and letter-to-sound linking. In the letter-name identification task, the teacher presents in random order the letters of the alphabet in uppercase and lowercase, asking for each letter's name. In the letter-to-sound linking task, the teacher first asks the child to isolate the first sound in a word ("lamp"→/l/). Then the teacher shows the child three letters (c, o, l) and asks the child to point to the letter that makes that sound. Stuart found this kind of letter-to-sound linking task to be more predictive of successful reading than Clay's screening battery.[78] There may be kindergartners who know letter names and can rhyme and blend onset-rhymes, but are not yet successful at linking letters to sounds. For them, instruction should focus on alliterative games and building words with a small set of taught letter-sounds.

In first grade the graphophonemic tasks require the student to write spelling patterns. The tasks progress from initial consonant substitution to final consonant substitution to medial

vowel substitution. For example, in the initial consonant substitution the teacher places a spelling pattern in front of the child (__ad) and spreads out seven consonant letters (d, f, h, m, p, s, t). Then the teacher asks the child to make the word "mad," followed by four additional words (dad, fad, tad, had). More difficult tasks require the substitution of initial and final blends (making "drip" from the pattern __ip and making "list" from the pattern li__). The second-grade inventory also includes four spelling tasks that cover long and complex vowel spellings, compound words, consonant digraphs, past tense, homophones, plurals, consonant doubling, and inflectional endings.

Comprehension tasks in kindergarten and first and second grades consist of narrative and expository passages drawn from children's books. During the 1998–1999 and 1999–2000 school years, passages from real books rather than artificially constructed passages were used in an attempt to provide an authentic performance measure of reading/listening comprehension. However, in our implementation study of 6,000 children in randomly selected schools in urban, suburban, and rural Texas, we found that the vast majority of first-graders could not read the authentic text said to be at first-grade level. Therefore, for the 2000–2001 edition of the TPRI we constructed first-grade passages that progress in difficulty across the year with respect to word properties of sound-spelling patterns and word frequency. Students are placed in passages at their instructional level based on their performance on a list of words linked empirically to oral reading accuracy in the passages. Instructional level is defined as the level at which oral reading accuracy ranges from 90 percent to 94 percent.

In administering the TPRI, the question is whether the graphophonemic knowledge assessed in the inventory and the attention and memory skills not assessed in the inventory—but required of any complex performance—will transfer to the comprehension tasks. In kindergarten, children listen to the teacher read a passage, then they answer questions. In first and second grades the children are asked to read the passage and to answer orally the questions the teacher asks. If a student miscalls more

than three words in the first sentence, then the teacher turns the task into a listening comprehension exercise. However, the new procedure of placing students in instructional-level text through performance on a word list minimizes the need for listening comprehension in grades one and two. As the student reads the passage aloud, the teacher notes miscalled words by slashing them in the student booklet. After five seconds have elapsed, the teacher provides the student with the word. When the student has finished reading the passage, the teacher can count the number of miscalled words and circle the reading accuracy level—frustrational (more than 10 percent), instructional (6 percent to 10 percent errors), or independent (less than 6 percent errors).

In the 2000–2001 edition of the TPRI, we added reading fluency rate. Teachers are provided with a stopwatch to time the students while they read a passage aloud. In order to calculate reading fluency rate, teachers are told to (a) determine the number of words read correctly and multiply by 60 and (b) divide this number by the number of seconds it took the student to read the passage. The reading rate goal is 60 WPM by the end of first grade and 90 WPM by the end of second grade.

Several passages are provided for the beginning, middle, and end of the year for first and second grades (with kindergarten having only middle and end-of-year passages) so that the teacher can note progress on increasingly complex texts. In the 2004–2005 edition, fluency probes are provided so that fluency may be monitored as often as twice a month for at-risk students. Passage complexity is determined empirically through an item development study, rather than by readability. Readability formulas are typically based on the number of words, syllables, and sentences in the text being evaluated. We found that the formulas produced highly variable results for beginning reading passages such as those used in the TPRI. The five questions that follow the TPRI passage to assess comprehension vary in the extent to which the answer is explicitly stated in the passage. Roughly, three of the five questions are explicit and two are implicit. Implicit questions require the student to make an inference about events, themes, or characters. Here, as an example of this procedure,

we provide one of the passages used for the beginning of the first grade in the first edition of the TPRI. The passage is from *Danny and the Dinosaur,* by Syd Hoff (and used by permission of HarperCollins Publishers):

> One day Danny went to the museum. He wanted to see what was inside. He saw Indians. He saw bears. He saw Eskimos. He saw guns. He saw swords. And he saw . . . DINOSAURS! Danny loved dinosaurs. He wished he had one. "I'm sorry they are not real," said Danny. "It would be nice to play with a dinosaur." "And I think it would be nice to play with you," said a voice.

After reading or listening to this passage, the child is asked to answer these questions:

1. Where did Danny go?

   Correct: To the museum

2. Tell me two things that Danny saw at the museum.

   Correct: Two of the following: Indians, bears, Eskimos, guns, swords, dinosaurs

3. What did Danny love most?

   Correct: Dinosaurs

4. What did Danny want to do with the dinosaurs?

   Correct: To play

5. What do you think talked to Danny? (Note: If the student answers "a voice," ask them whose voice.)

   Correct: A dinosaur

These five questions serve as a probed retelling. Full story retelling places demands on discourse skills that may interfere with assessment of reading comprehension. However, as Morrow points out, "Retelling allows the child to reconstruct meaning and personalize information."[79]

To investigate whether story retelling contributes to sense of story above and beyond the five questions in the TPRI and to examine the effect of reading the passage in the TPRI booklet, in the actual storybook, or in a guided reading context, we con-

ducted a small study with 124 first-graders as part of the validation study.[80] Each child read two short passages taken from children's literature books, answered five comprehension questions, and retold the story they had just read. Reading rate and accuracy were recorded for each child for each passage. There were no differential effects of context on reading comprehension as measured by comprehension questions or retell scores. That is, reading a storybook passage printed in the TPRI booklet versus reading the passage in the storybook with or without adult scaffolding made no difference in the number of story grammar elements in first-graders' retells or number of comprehension questions answered correctly. For one passage—the one that was above a first-grade level—there were significant effects of fluency (that is, reading rate) on answering comprehension questions and on retell scores. Specifically, speed and accuracy of decoding explained 13 percent of the variance in correctly answering comprehension questions and 10 percent of the variance in retelling scores. Moderate correlations were found among formal and informal measures of reading comprehension.

Validation of the TPRI

The TPRI screen provides extensive psychometric data. The screen is based on empirically based predictors of reading success at the end of grades one and two. These predictors were derived from a study that had a modified, longitudinal, time-sequential design in which 945 children in kindergarten and first and second grades were evaluated on reading and reading-related skills four times yearly for one to three years. In addition, achievement tests were administered at the end of first and second grades. The participating children were in regular education in three elementary schools. The percentage of participation in the federal lunch program at the three schools was 13 percent, 15 percent, and 30 percent. The student populations varied in socioeconomic status from lower-middle to upper-middle class. The sample was approximately half boys and half girls. The ethnic composition of the sample was diverse, with the following breakdown in kindergarten: 54 percent Caucasian, 18 percent African-American,

15 percent Hispanic, 12 percent Asian, and 1 percent other. Children were excluded from the sample if they had severe emotional problems, uncorrected vision problems, hearing loss, or acquired neurological disorders or were classified at the lowest level of English as a second language (ESL). Children who were at ESL levels 2, 3, and 4 were included in the sample.

The items on the screen were those items selected on the basis of Item Response Theory (IRT) from a larger battery of items that discriminate success and failure on reading outcomes at the end of grades one and two.[81] The larger battery included measures of visual-motor integration, visual-spatial skill, expressive and receptive syntax, phonological memory, vocabulary, attention, IQ, rapid naming, letter names and letter sounds knowledge, phonological awareness, word reading, and spelling. For kindergarten, we attempted to predict outcomes using the Woodcock-Johnson PsychoEducational Test Battery Basic Reading cluster. For predictions involving first- and second-graders, the Woodcock-Johnson Broad Reading cluster, which consists of letter-word identification and word attack measures and a cloze-based passage comprehension measure, was used. The criteria for risk were arbitrarily set at grade equivalents of 1.4 or lower at the end of grade one and 2.4 or lower at the end of grade two on the Woodcock-Johnson. In first grade this grade equivalent represents the 22nd percentile for Basic Reading and the 18th percentile for Broad Reading. In second grade it represents the 35th percentile. The cut-point was deliberately set higher in grade two because of the greater stability in the prediction equations and the reduction in time available for a student to reach the Texas Reading Initiative goal of being on or above grade level by the end of third grade.

Separate analyses were conducted on the five assessment timepoints. We attempted to establish a series of prediction equations that helped select variables contributing uniquely to the prediction. Decisions about effective predictors were based on both the accuracy of individual child predictions and the relation of false positive and false negative errors. The goal in selecting the best prediction set was to maximize identification, minimize the num-

ber of predictors, and produce the lowest possible false positive errors rate, keeping false negative error rates below 10 percent. False positives and false negatives are inherent to any assessment device and are inevitably linked. To use screening as an example, false negatives occur when a child meets criteria on the screening but fails to learn to read; a false positive occurs when a child does not meet criteria on the screening but nevertheless becomes a successful reader. A false negative error is more serious because these children do not receive the additional assistance they require at the earliest possible time, which makes their problems more difficult to remediate at a later time. False positive errors are a concern because they place an increased demand on scarce resources. False positives in kindergarten and the beginning of first grade may reflect the assessment of children from poor neighborhoods or who have limited English proficiency whose opportunity to become literate comes from instruction at school. False positive rates in kindergarten ranged from 44 percent in December to 38 percent in April, to 36 percent in first grade, and to less than 15 percent at the beginning of second grade.[82]

In order to collect reliability and validity data for the TPRI inventory, a field study was conducted in four Houston Independent School District elementary schools that were also participating in a much larger study of early reading funded by the National Institute of Child Health and Human Development (NICHD). The field study involved thirty-two kindergarten and first-grade teachers, 128 kindergarten students, and 144 first-grade students. In each classroom, eight students were randomly selected from the sample of all NICHD students to participate in the field study. We trained the teachers to administer the TPRI, then provided substitutes so that these teachers could administer the TPRI screen and inventory to their own students on one day and to the students in the neighboring classroom on the next day. We included this step to obtain inter-rater of scoring between a teacher who knows the student well and another teacher of the same grade who does not know the student well.

Included in the analyses of the field study data were (a) concurrent validation of the TPRI with well-known individually

administered word recognition and comprehension measures (that is, the letter-word identification and passage comprehension from the Woodcock-Johnson),[83] (b) internal consistency of items, and (c) inter-rater reliability judgments. These judgments were obtained in three ways. First, to examine scoring accuracy, a teacher's ability to apply scoring criteria was compared with an expert's scoring of the same protocol. Second, to examine objectivity of scoring, teachers administered the TPRI to students who were either from their own classroom or from the neighboring teacher's classroom. Third, to see if they agreed on interpretation of results, teachers were asked to rate the importance of various instructional strategies for an individual student, based on that student's TPRI results, then to prioritize those strategies on one week's lesson plans.

The evidence for reliability for the items in the screen and inventory was very good. Median internal consistencies were .89 for the end of kindergarten, .80 for the beginning of first grade, .74 for the end of first grade, and .68 for the beginning of second grade. The median lower-bound estimate of test-retest reliability was .60. Only the subtest for book and print awareness had unacceptably low internal consistency, test-retest reliability, and inter-rater reliability. Evidence for the TPRI's construct validity was provided through correlations with the Woodcock-Johnson (WJ) reading scores and scores from the Gray Oral Reading Test-III (GORT-III). In second grade, the correlations with the WJ ranged between .26 and .61, and correlations with the GORT-III ranged between .23 and .56, depending on the TPRI passage. In first grade, the correlation with WJ was .48 and the correlations with GORT-III comprehension, reading rate, and reading accuracy ranged from .41 to .52 on the passage with adequate reliability. The passage from *Danny and the Dinosaur* had inadequate reliability because the children were very familiar with the story. In kindergarten, there were no additional measures of reading comprehension. The strongest correlation with the two passages in the TPRI listening comprehension subtest was provided by the Peabody Picture Vocabulary Test-Revised (PPVT-R). Correlations ranged from .41 to .65 in kindergarten. In first grade, the correla-

tions of TPRI reading comprehension with the PPVT-R ranged from .32 to .38; in second grade, the range was .53 to .63 (see the TPRI technical manual for details).

Teachers participating in this field study also provided their opinions about training issues and test administration issues. Overall, teachers responded positively to the presentation format and informational content of the TPRI training. At least 70 percent found the directions and the format of the teacher's guide and student booklet clear. The majority of teachers felt that all parts of the TPRI were easy to administer and were useful. They rated the TPRI very helpful for identifying strengths and weaknesses of students not previously taught and were very likely to recommend the TPRI to another teacher or administrator. Most teachers responded that gathering materials for administration of the TPRI and planning individual instruction based on TPRI results would be relatively easy. Forty percent were familiar with other reading assessments and reported that when comparing the TPRI with the other assessments, the TPRI was better in terms of ease of administration, usefulness for planning instruction, identification of students' reading strengths and weaknesses, and worthwhile use of instructional time. The majority of teachers did not think that changes should be made to either the screening or the inventory portions of the instrument.

Professional Development

If teachers rather than testing professionals are to administer the assessment, then professional development is necessary. Typically, teacher certification does not require coursework in assessment, diagnosis, and intervention. These courses are more commonly found at the master's degree level. The areas of the TPRI in which teachers require the most training the TPRI are: (a) pronunciation of speech sounds in the phonemic awareness and letter-sound tasks; (b) learning to assess rather than to coach; and (3) developing intervention strategies based on the results of assessment. Teacher certification typically does not include information about phonology. Therefore, something as seemingly simple as pronouncing letter-sounds requires professional development. For

example, it should be pointed out that the letter *p* represents /p/, not "puh;" the letter *m* represents /m/ or /mmmmm/, but not "muh."

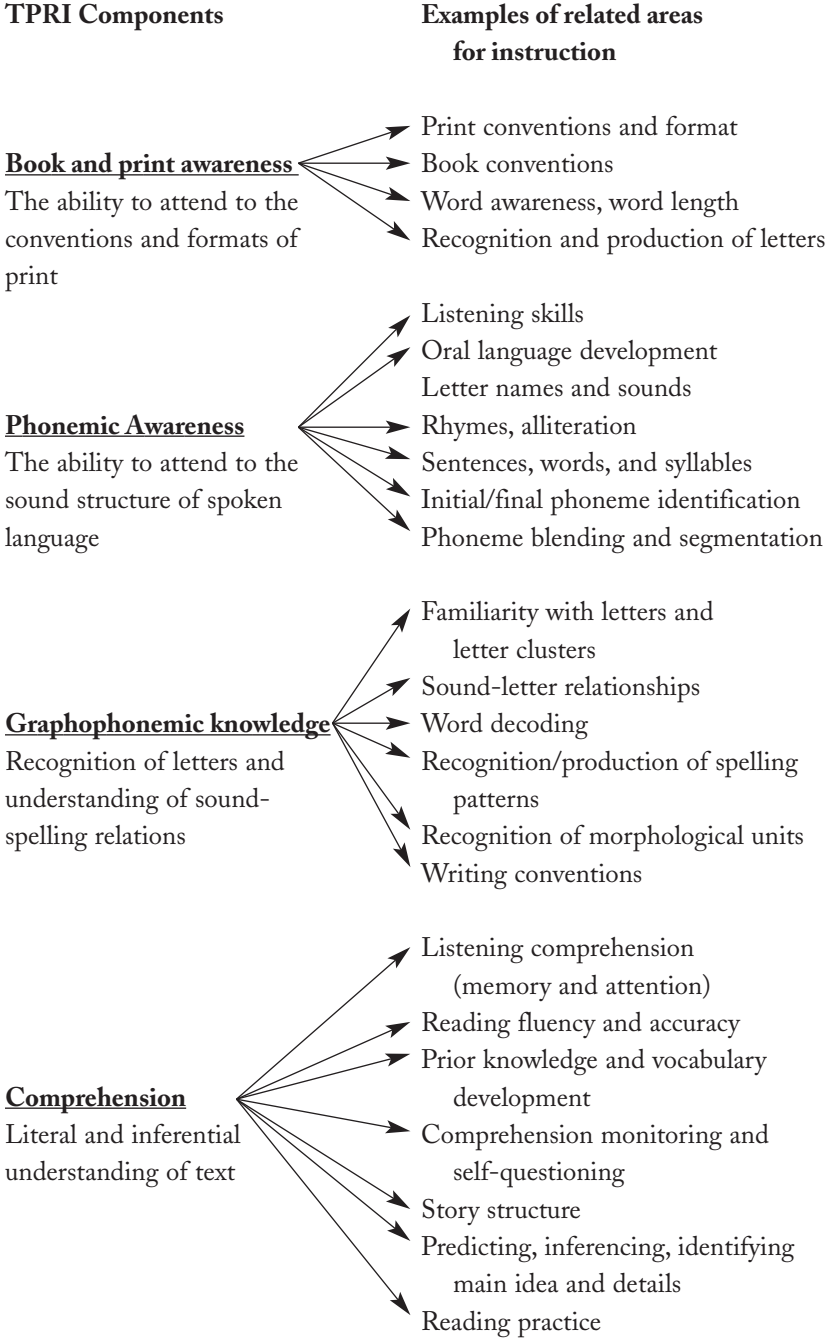
Coaching occurs because teachers are used to teaching rather than assessing and also because many teachers feel that assessment and instruction should occur concurrently. However, there is a time for teachers to step back from classroom instruction and put on the assessor's hat to see if knowledge and skills transfer to new contexts. If teachers are not willing to do this, assessment might be taken out of their hands. This would be unfortunate because it is teachers who are in the best position to use results of assessment to affect instruction.

To help with intervention strategies, we have developed an Intervention Activities Guide as part of the TPRI kit that links results from the TPRI with specific classroom activities. But the first line of defense is prevention. As teachers in Texas align their curriculum with the adopted state curriculum standards (the Texas Essential Knowledge and Skills [TEKS]), they will find the components of the TPRI inventory closely aligned with TEKS objectives. Examples of the link between curricular standards and the TPRI are provided in Figure 3.2.

Linking TPRI to instruction

The TEKS and the TPRI go hand in hand. As Simmons and Resnick point out,[84] "without performance standards the meaning of content standards is subject to interpretations" (12). Thus, performance on the TPRI inventory provides a concrete demonstration of the knowledge and skills supposedly covered in the classroom curriculum. By readministering the inventory at midyear and end-of-year, teachers can determine progress toward the standards. Because the items on the kindergarten and first- and second-grade screens predict that students will be on or above grade level at the end of first and second grades, teachers have a gauge with which to calibrate benchmark expectations from year to year. Thus, no child should fall through the cracks because he or she appears to be making sufficient progress on the inventory. The end-of-year screen in kindergarten and first

FIGURE 3.2 TPRI inventory and related instruction areas



grade and the beginning-of-year screen in first and second grades provide safeguards. In fact, the identification of risk is so accurate at the beginning of second grade (above 85 percent) that further evaluation is warranted. That student will be well below grade level unless intervention is undertaken.

Conclusions

We have argued for the importance of assessing early reading skills by pointing out (a) the intractable nature of reading problems identified in third grade and beyond, (b) the presence of risk characteristics in kindergarten and grade one, and (c) the effectiveness of early intervention. The major impediments to assessing early reading skills are (a) the "wait and see" adage or "late bloomer" attribution espoused by some early childhood educators and (b) the need to accumulate sufficient failure on standardized achievement tests before an IQ test is administered to determine eligibility for special education. We have described six early assessment systems. Three of these were local efforts that engage teachers in collecting evidence upon which to base individual curricular decisions. Three were large-scale efforts—the California Learning Record, the Work Sampling System, and the Phonological Awareness and Literacy Screening currently under development in Virginia. Both the Work Sampling System and PALS have demonstrated evidence of validity and reliability. Finally, we have presented the Texas Primary Reading Inventory in detail as an example of a large-scale statewide assessment aligned to state curriculum standards and based on psychometric evidence that includes predictive validity. The items on the short screening component of the TPRI allow a teacher to know quickly which students are on track to becoming successful readers one year and two years later. That prediction allows the teacher to administer the more time-consuming inventory to the students potentially at risk so that instructional objectives can be established and monitored for progress.

Because the false positive rate is relatively large in kindergarten and first grade, early reading assessment should *not* be part of the

accountability system. In other words, many kindergartners and first-graders will appear to be at risk for reading failure when, in fact, they turn out to become successful readers. However, by the beginning of second grade, the false positive rate is below 15 percent. Therefore, a second-grader who does not meet the criterion on the TPRI screen is a candidate for further evaluation and intervention to avoid being well below grade level at the end of the year. Thus, the performance on the Texas Primary Reading Inventory not only signals the need for early intervention, it holds the promise of preventing reading difficulties from occurring by maximizing the individual student's opportunities to learn in the classroom.

Assessment of early reading skills is useful only to the extent that (a) assessment results can be put in the hands of the teacher and used to plan instructional objectives and (b) the results enhance instructional outcomes by providing information to parents, teachers, administrators, and policy makers on the efficacy of different programs and decisions. Here Linn and colleagues and Pearson's notion of consequential validity is important.[85] Assessment for the sake of assessment is not meaningful. In the area of reading, assessment decisions must be linked to decision-making processes that will enhance reading outcomes for children. Decisions to assess early reading skills should be linked to the teacher's ability to plan instructional objectives. The use of assessments for accountability must not be punitive, but linked to the statewide curriculum and goals for all participants in the educational community. In the area of early reading assessments, the goal should be prevention so that accountability goals can be met later in schooling. This is consistent with the Reading First component of the Elementary and Secondary Education Act's No Child Left Behind legislation that requires assessment in K–3 classrooms for screening, diagnosis, progress monitoring, and outcome.

Notes

1. Sally E. Shaywitz, "Dyslexia," *Scientific American* 275 (1996): 98–104.
2. Catherine E. Snow, M. Susan Burns, and Patricia Griffin, *Preventing Reading Difficulties in Young Children* (Washington, D.C.: National Academy of Science, 1998).

3. Ibid.
4. Ibid.
5. David J. Francis, Sally E. Shaywitz, Karla K. Stuebing, Bennett A. Shaywitz, and Jack M. Fletcher, "Developmental Lag Versus Deficit Models of Reading Disability: A Longitudinal Individual Growth Curves Analysis," *Journal of Educational Psychology* 88 (1996): 3–17.
6. Connie Juel, "Learning to Read and Write: A Longitudinal Study of 54 Children from First Through Fourth Grades," *Journal of Educational Psychology* 80 (1988): 437–47.
7. Linda M. Phillips, Stephen P. Norris, Wendy C. Osmond, and Agnes M. Maynard, "Relative Reading Achievement: A Longitudinal Study of 187 Children from First Through Sixth Grades," *Journal of Educational Psychology* 94 (2002): 3–13.
8. Joseph K. Torgesen, "The Prevention and Remediation of Reading Disabilities: Evaluating What We Know from Research," *Journal of Academic Language Therapy* 1 (1997): 11–47.
9. Joseph K. Torgesen, Richard K. Wagner, Carol A. Rashotte, Elaine Rose, Patricia Lindamood, Tim Conway, and Cyndi Garvan, "Preventing Reading Failure in Young Children with Phonological Processing Disabilities: Group and Individual Responses to Instruction," *Journal of Educational Psychology* 91 (1999): 579–93; Joseph K. Torgesen, A. W. Alexander, Richard K. Wagner, Carol A. Rashotte, K. S. Voeller, and Tim Conway, "Intensive Remedial Instruction for Children with Severe Reading Disabilities: Immediate and Long-Term Outcomes from Two Instructional Approaches," *Journal of Learning Disabilities* 34 (2001): 33–58.
10. Frank R. Vellutino, Donna M. Scanlon, Edward Sipay, Sheila Small, Alice Pratt, Rusan Chen, and Martha Denckla, "Cognitive Profiles of Difficult-to-Remediate and Readily Remediated Poor Readers: Early Intervention As a Vehicle for Distinguishing Between Cognitive and Experimental Deficits As Basic Causes of Specific Reading Disability," *Journal of Educational Psychology* 88 (1996): 601–38.
11. Sandra Iversen and William Tunmer, "Phonological Processing Skills and the Reading Recovery Program," *Journal of Educational Psychology* 85 (1993): 112–26; William E. Tunmer, James W. Chapman, Henry Ryan, and Jane E. Prochnow, "The Importance of Providing Beginning Readers with Explicit Training in Phonological Processing Skills," *Australian Journal of Learning Disabilities* 3 (1998): 4–14 (hereafter cited as "The Importance of Providing Training in Phonological Processing Skills").
12. Barbara R. Foorman, David J. Francis, Jack M. Fletcher, Chris Schatschneider,

and Paris Mehta, "The Role of Instruction in Learning to Read: Preventing Reading Failure in At-Risk Children," *Journal of Educational Psychology* 90 (1998): 38–55.

13. Gerald Bond and Miles A. Tinker, *Reading Difficulties: Their Diagnosis and Correction* (New York: Appleton-Century-Crofts, 1957), 82.
14. Ibid., 70.
15. Michael Rutter and William Yule, "The Concept of Specific Reading Retardation," *Journal of Child Psychology and Psychiatry* 16 (1975): 181–97.
16. Jack M. Fletcher, Sally E. Shaywitz, Donald P. Shankweiler, Leonard Katz, Isabelle Y. Liberman, Ann Fowler, David J. Francis, Karla K. Stuebing, and Bennett A. Shaywitz, "Cognitive Profiles of Reading Disability: Comparisons of Discrepancy and Low Achievement Definitions," *Journal of Educational Psychology* 85 (1994): 1–18; Jack M. Fletcher, David J. Francis, Sally E. Shaywitz, Barbara R. Foorman, Karla K. Stuebing, and Bennett A. Shaywitz, "Intelligent Testing and the Discrepancy Model for Children with Learning Disabilities," *Learning Disabilities Research and Practice* 13 (1998): 186–203; Keith E. Stanovich and Linda S. Siegel, "Phenotypic Performance Profiles of Children with Reading Disabilities: A Regression-Based Test of the Phonological-Core Variable Difference Model," *Journal of Educational Psychology* 86 (1994): 24–53; Karla K. Stuebing, Jack M. Fletcher, J. M. LeDoux, G. Reid Lyon, Sally Shaywitz, and Bennett Shaywitz, "Validity of IQ-Discrepancy Classifications of Reading Disabilities: A Meta-Analysis," *American Educational Research Journal* 39 (2001): 469–518.
17. Cecil R. Reynolds, "Critical Measurement Issues in Learning Disabilities," *Journal of Special Education* 18 (1984): 451–76.
18. Rutter and Yule, "The Concept of Specific Reading Retardation."
19. Jack M. Fletcher, G. Reid Lyon, Marcia Barnes, Karla K. Stuebing, David Francis, Richard K. Olson, Sally E. Shaywitz, and Bennett A. Shaywitz, "Classification of Learning Disabilities: An Evidenced-Based Evaluation" in *Identification of Learning Disabilities: Research to Practice,* ed. R. Bradley, L. Danielson, and D. Hallahan (Mahwah, N.Y.: Erlbaum, 2002): 185–250; G. Reid Lyon, Dwayne Alexander, and Stephen Yaffe, "Progress and Promise in Research on Learning Disabilities," *Learning Disabilities* 8 (1997): 1–6; G. Reid Lyon, Jack M. Fletcher, Sally E. Shaywitz, Bennett A. Shaywitz, Joseph K. Torgesen, Frank B. Wood, Ann Schulte, and Richard Olson, "Rethinking Learning Disabilities" in *Rethinking Special Education for a New Century,* ed. C. E. Finn Jr., R. A. J. Rotherham, and C. R. Hokanson Jr. (Washington, D.C.: Thomas B. Fordham Foundation and

Progressive Policy Institute, 2001): 259–87; David L. Share, Robert McGee, and Philip D. Silva, "I.Q. and Reading Progress: A Test of the Capacity Notion of I.Q.," *Journal of the American Academy of Child and Adolescent Psychiatry* 28 (1989): 97–100; Stuebing, Fletcher, et al., "Validity of IQ-Discrepancy Classifications"; Joseph K. Torgesen, and Richard K.Wagner, "Alternative Diagnostic Approaches for Specific Developmental Reading Disabilities," *Learning Disabilities Research & Practice* 13 (1998): 220–32.

20. Barbara R. Foorman, "The Relevance of a Connectionistic Model of Reading for 'The Great Debate,'" *Educational Psychology Review* 6 (1994): 25–47; Michael W. Harm and Mark S. Seidenberg, "Phonology, Reading Acquisition, and Dyslexia: Insights from Connectionist Models," *Psychological Review* 106 (1999): 491–528; David C. Plaut, James L. McClelland, Mark S. Seidenberg, and Kathleen Patterson, "Understanding Normal and Impaired Word Reading: Computational Principles in Quasi-Regular Domains," *Psychological Review* 103 (1996): 56–115; Keith Rayner, Barbara Foorman, Charles A. Perfetti, David Pesetsky, and Mark S. Seidenberg, "How Should Reading Be Taught?" *Scientific American* (2002): 84–91; Stanovich and Siegel, "Phenotypic Performance Profiles."
21. Jack M. Fletcher and G. Reid Lyon, "Reading: A Research-Based Approach" in *What's Gone Wrong in America's Classrooms,* ed. W. M. Evers (Stanford, Calif.: Hoover Institution Press, 1998), 49–90.
22. Shaywitz, "Dyslexia"; Keith E. Stanovich, "Cognitive Science Meets Beginning Reading," *Psychological Science* 2 (1991): 70–81; Frank R. Vellutino, "Introduction to Three Studies on Reading Acquisition: Convergent Findings on Theoretical Foundations of Code-Oriented Versus Whole-Language Approaches to Reading Instruction," *Journal of Educational Psychology* 83 (1991): 437–43.
23. David L. Share and Keith E. Stanovich, "Cognitive Processes in Early Reading Development: A Model of Acquisition and Individual Differences," *Issues in Education: Contributions from Educational Psychology* 1 (1995): 1–57.
24. Sally E. Shaywitz, Michael D. Escobar, Bennett A. Shaywitz, Jack M. Fletcher, and Robert Makuch, "Distribution and Temporal Stability of Dyslexia in an Epidemiological Sample of 414 Children Followed Longitudinally," *New England Journal of Medicine* 326 (1992): 145–50.
25. Sally E. Shaywitz, Bennett A. Shaywitz, Jack M. Fletcher, and Michael D. Escobar, "Prevalence of Reading Disability in Boys and Girls: Results of the Connecticut Longitudinal Study," *Journal of the American Medical Association* 264 (1990): 998–1002.
26. Magdalen D. Vernon, *Backwardness in Reading: A Study of Its Nature and Origin* (London: Cambridge University Press, 1958).

27. Shaywitz, Shaywitz, et al., "Prevalence of Reading Disability in Boys and Girls."
28. Lyon, Fletcher, et al., eds., "Rethinking Learning Disabilities"; Torgesen and Wagner, "Alternative Diagnostic Approaches"; Share, McGee, and Silva, "I.Q. and Reading Progress."
29. Fletcher, Lyon, et al., *Classification of Learning Disabilities;* Fletcher, Shaywitz, et al., "Cognitive Profiles;" Stuebing, Fletcher, et al., "Validity of IQ-Discrepancy Classifications;" Stanovich and Siegel, "Phenotypic Performance Profiles."
30. G. Reid Lyon and Judith Rumsey, eds., *Neuroimaging: A Window to the Neurological Foundations of Learning and Behavior in Children* (Baltimore: Paul C. Brookes, 1997). Sally E. Shaywitz, Bennett A. Shaywitz, Ken R. Pugh, Robert K. Fulbright, Robert T. Constable, William E. Mencl, Donald P. Shankweiler, Alvin M. Liberman, Pawel Skudlarski, Jack M. Fletcher, Leonard Katz, Karen E. Marchione, Cheryl Lacadie, Carol Gatenby, and John C. Gore, "Functional Disruption in the Organization of the Brain for Reading in Dyslexia," *Proceedings of the National Academy of Sciences* 95 (1998): 2636–41. Panagiotis G. Simos, Jack M. Fletcher, Barbara R. Foorman, David J. Francis, E. M. Castillo, M. F. Davis, Patricia G. Mathes, Carolyn A. Denton, and Andrew C. Papanicolaou, "Brain Activation Profiles During the Early Stages of Reading Acquisition," *Journal of Child Neurology* 17 (2002): 159–63.
31. Jack M. Fletcher, Bonnie Brookshire, Timothy P. Bohan, Michael Brandt, and Kevin Davidson, "Early Hydrocephalus" in *Syndrome of Nonverbal Learning Disabilities: Neurodevelopmental Manifestations,* ed. B. P. Rourke (New York: Guilford Press, 1995), 206–38.
32. Lon Cardon, S. D. Smith, David Fulker, B. S. Kimberling, Bruce Pennington, and J. C. DeFries, "Quantitative Trait Locus for Reading Disability on Chromosome 6," *Science* 226 (1994): 276–79. Elena L. Grigorenko, Frank B. Wood, Marianne S. Meyer, Lesley A. Hart, William C. Speed, Arlene Shuster, and Donna Pauls, "Susceptibility Loci for Distinct Components of Developmental Dyslexia on Chromosomes 6 and 15," *American Journal of Human Genetics* (1997).
33. Catherine E. Snow, M. Susan Burns, and Patricia Griffin, *Preventing Reading Difficulties in Young Children* (Washington, D.C.: National Academy of Sciences, 1998).
34. Foorman, Francis, et al., "The Role of Instruction." Maureen W. Lovett, Lea Lacerenza, Susan L. Borden, Jan C. Frijters, Karen A. Steinbach, and Maria De Palma, "Components of Effective Remediation for Developmental Reading Disabilities: Combining Phonological and Strategy-Based Instruction," *Journal of Educational Psychology* 92 (2000): 263–83. Torgesen, "Prevention and Remediation." Torgesen, Wagner, et al., "Preventing Reading Failure." Torgesen,

Alexander, et al., "Intensive Remedial Instruction." Vellutino, Scanlon, et al., "Cognitive Profiles."

35. Foorman, Francis, et al., "The Role of Instruction in Learning to Read." Lovett, Lacerenza, et al., "Components of Effective Remediation." Torgesen, "Prevention and Remediation." Torgesen, Wagner, et al., "Preventing Reading Failure." Torgesen, Alexander, et al., "Intensive Remedial Instruction." Vellutino, Scanlon, et al., "Cognitive Profiles."
36. J. Bruce Tomblin and Xuyang Zhang, "Language Patterns and Etiology in Children with Specific Language Impairment" in *Neurodevelopmental Disorders: Contributions to a New Framework,* ed. H. Tager-Flusberg (Cambridge, Mass.: MIT Press, 1999), 361–82.
37. Frank R. Vellutino, Dyslexia: *Theory and Research* (Cambridge, Mass.: MIT Press, 1979).
38. Keith E. Stanovich, "Romance and Reality," *The Reading Teacher* 47 (1994): 280–9.
39. Foorman, Francis, et al., "The Role of Instruction in Learning to Read." Torgesen, "Prevention and Remediation." Torgesen, Wagner, et al., "Preventing Reading Failure." Torgesen, Alexander, et al., "Intensive Remedial Instruction."
40. Lorie Shepard, "What Policy Makers Who Mandate Tests Should Know About the New Psychology of Intellectual Ability and Learning" in *Changing Assessments: Alternative Views of Aptitude, Achievement and Instruction,* ed. B. R. Gifford and M. C. O'Connor (Boston, Mass.: Kluwer Academic Publishers, 1992), 301–28.
41. Jerome Sattler, *Assessment of Children,* 3rd ed. (San Diego, Calif.: Author, 1992).
42. Anne Anastasi, *Psychological Testing,* 5th ed. (New York: Macmillan, 1982).
43. Richard W. Woodcock, Kevin S. McGrew, and Nancy Mather, *Woodcock-Johnson* 3 (Itasca, Ill.: Riverside Publishing, 2001).
44. Lauren B. Resnick and Daniel P. Resnick, "Assessing the Thinking Curriculum: New Tools for Educational Reform" in *Changing Assessments,* ed. Gifford and O'Connor, 37–76.
45. Lynn S. Fuchs and Doug Fuchs, "Treatment Validity: A Unifying Concept for Reconceptualizing Identification of Learning Disabilities," *Learning Disabilities Research and Practice* 13 (1998): 45–58.
46. William A. Mehrens, "Using Performance Assessment for Accountability Purposes," *Educational Measurement: Issues and Practice* (spring 1992): 3–9. Reprinted in this volume.
47. Lauren Leslie and Joanne Caldwell, *Qualitative Reading Inventory* 2 (New York:

HarperCollins, 1995). Leslie M. Morrow, Michael Pressley, Jeffery K. Smith, and Michael Smith, "The Effect of a Literature-Based Program Integrated into Literacy and Science Instruction with Children from Diverse Backgrounds," *Reading Research Quarterly* 32 (1997): 54–77.

48. Robert L. Linn, Eve L. Baker, and Samuel B. Dunbar, "Complex Performance-Based Assessment: Expectations and Validation Criteria," *Educational Leadership* (1991): 15–21.
49. Elfrieda H. Hiebert, Sheila W. Valencia, and Peter P. Afflerbach, "Definitions and Perspectives." in *Authentic Reading Assessment: Practices and Possibilities,* ed. S. W. Valencia, E. H. Hiebert, and P. P. Afflerbach (Newark, Del.: International Reading Association, 1994), 6–21. Grant Wiggins, "Teaching to the Authentic Test," *Educational Leadership* 45 (1989).
50. Marie M. Clay, *Reading Recovery: A Guidebook for Teachers in Training* (Portsmouth, N.H.: Heinemann, 1993). Marie M. Clay, *The Early Detection of Reading Difficulties,* 3rd ed. (Portsmouth, N.H.: Heinemann, 1985).
51. Clay, *Reading Recovery: A Guidebook.*
52. Gay Sue Pinnell, Carol A. Lyons, Diane E. DeFord, Anthony S. Bryk, and Melvin Selzer, "Comparing Instructional Models for the Literacy Education of High-Risk First-Graders," *Reading Research Quarterly* 29 (1994): 8–38.
53. Tom Nicholson, "A Comment on Reading Recovery," *New Zealand Journal of Educational Studies* 24 (1989): 95–97. Valerie Robinson, "Some Limitations of Systemic Adaptation: The Implementation of Reading Recovery," *New Zealand Journal of Educational Studies* 24 (1989): 35–45.
54. Iversen and Tunmer, "Phonological Processing Skills."
55. Yola K. Center, Kevin Wheldall, Louella Freeman, Lynne Outhred, et al., "An Evaluation of Reading Recovery," *Reading Research Quarterly* 30 (1995): 240–63.
56. Tunmer, Chapman, et al., "The Importance of Providing Training in Phonological Processing Skills."
57. Snow, M. Burns, and Griffin, *Preventing Reading Difficulties.*
58. Terry Salinger and Edward Chittenden, "Analysis of an Early Literacy Portfolio: Consequences for Instruction," *Language Arts* 71 (1994): 446–52.
59. Ibid.
60. Mary A. Barr, S. Ellis, H. Tester, and A. Thomas, *The Primary Language Record: Handbook for Teachers* (Portsmouth, N.H.: Heinemann, 1988). Beverly Falk, "Using Direct Evidence to Assess Student Progress: How the Primary Language Record Supports Teaching and Learning." in *Assessing Reading 1: Theory and Practice,* ed. C. Harrison and T. Salinger (New York: Routledge, 1998), 152–65.

61. Mary A. Barr, ed., *California Learning Record: Handbook for Teachers* (San Diego: Center for Language in Learning, 1995).
62. Mary A. Barr and P. J. Hallam, *California Learning Record,* 1996 Moderation Report (San Diego: Center for Language in Learning, 1996).
63. James Hoffman, Nancy Roser, and Jo Worthy, "Challenging the Assessment Context for Literacy Instruction in First Grade: A Collaborative Study" in Assessing *Reading* 1, ed. Harrison and Salinger: 166–81.
64. Donna M. Ogle, "K-W-L: A Teaching Model That Develops Active Reading of Expository Text," *The Reading Teacher* 39 (1986): 564–70.
65. Hoffman, Roser, and Worthy, "Challenging the Assessment Context."
66. Ibid.
67. P. David Pearson, "Standards and Assessment: Tools for Crafting Effective Instruction?" in *Literacy for All: Issues in Teaching and Learning,* ed. F. Lehr and J. Osborn (New York: Guilford Publications, Inc., 1998): 264–88.
68. Samuel J. Meisels, "Using Work Sampling in Authentic Assessments," *Educational Leadership* (winter 1996–97): 60–65.
69. Samuel J. Meisels, F. Liaw, Anthony Dorfman, and Robert Nelson, "The Work Sampling System: Reliability and Validity of a Performance Assessment for Young Children," *Early Childhood Research Quarterly* 10 (1995): 277–96.
70. Pearson, "Standards and Assessment."
71. Ibid.
72. Ibid.
73. *Texas Primary Reading Inventory Technical Manual* (Austin, Tex.: Texas Education Agency, 1998). (Available at www.trpi.org under Researchers/Psychometrics.)
74. Robert J. Tierney, "Literacy Assessment Reform: Shifting Beliefs, Principled Possibilities, and Emerging Practices," *The Reading Teacher* 51 (1998): 374–91.
75. *Texas Primary Reading Inventory Technical Manual.* Chris Schatschneider, Jack Fletcher, David Francis, Coleen Carlson, and Barbara Foorman, "Kindergarten Prediction of Reading Skills: A Longitudinal Comparative Study," *Journal of Educational Psychology* (in press). Torgesen, "Prevention and Remediation." Vellutino, Scanlon, Edward Sipay, et al., "Cognitive Profiles."
76. Clay, *Reading Recovery: A Guidebook.*
77. *Texas Primary Reading Inventory Technical Manual.*
78. Morag Stuart, "Prediction and Qualitative Assessment of Five- and Six-Year-Old Children's Reading: A Longitudinal Study," *British Journal of Educational Psychology* 65 (1995): 287–96. Marie M. Clay, *Reading Recovery: A Guidebook.*
79. Leslie M. Morrow, "Assessing Children's Understanding of Story Through Their

Construction and Reconstruction of Narrative" in *Assessment for Instruction in Early Literacy,* ed. L. M. Morrow and J. K. Smith (Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1990), 110–34.

80. Patricia McEnery, *The Role of Context in Comprehension of Narrative Text in First-Graders.* Unpublished doctoral dissertation (University of Houston, Houston, Tex., 1998).
81. Schatschneider, Fletcher, et al., *Kindergarten Prediction of Reading Skills.*
82. Jack Fletcher, Barbara Foorman, Amy Boudousquie, Marcia Barnes, Chris Schatschneider, and David Francis, "Assessment of Reading and Learning Disabilities: A Research-Based Intervention-Oriented Approach," *Journal of School Psychology* 40 (2002):27–63. Schatschneider, Fletcher, et al., *Kindergarten Prediction of Reading Skills.*
83. Richard W. Woodcock and M. Bonner Johnson, *Woodcock-Johnson Psychoeducational Battery* 3 (Allen, Tex.: DLM Teaching Resources, 2001).
84. Warren Simmons and Lauren Resnick, "Assessment As the Catalyst of School Reform," *Educational Leadership* (February 1993): 11–15.
85. Linn, Baker, and Dunbar, "Complex Performance-Based Assessment." Pearson, "Standards and Assessment."

Chapter 4

Science and Math Testing: What's Right and Wrong with the NAEP and the TIMSS?

Stan Metzenberg

Consider the question:

In the human body the digestion of proteins takes place primarily in which two organs?

A) Mouth and stomach

B) Stomach and small intestine

C) Liver and gall bladder

D) Pancreas and large intestine

The correct answer is B, and 66 percent of U.S. students in eighth grade answered it correctly on the National Assessment of Educational Progress (NAEP) Science Test administered in 2000.[1] It also is a very good question. It is stated in plain language

and has a single correct answer. The distracters (incorrect answers) would be plausible to a person who has not learned the material well.

Perhaps more important, the content being tested is a foundation for future study. In high school, these sixty-six out of 100 students can go on to learn how proteins in the food we eat are broken into amino acids by the action of enzymes, such as pepsin (in the stomach) and trypsin (in the intestine). They can learn how the release of these enzymes is regulated, how each enzyme works in different conditions of acidity, and how the amino acids released during digestion are absorbed into the blood vessels in the intestinal walls. It's a truly beautiful system.

The reader may think it vain for a biologist to wax poetic over the details of digestive physiology, as these matters do not weigh heavily on the minds of most adults. Can there really be a credible link between testing eighth-grade students on their knowledge of the small intestine and international competitiveness? Questions of this type are usually delivered with a smirk and by the outcomes-based educator. A century ago, it would have been, "Does the future laborer really need to memorize Latin clauses for later regurgitation? *Tu quidem non es qui hoc crederes!*" Educational policy is regularly mauled by this *reductio ad absurdum* argument, and building a stronger line of defense ought to be a key goal of reformers. The rational response is as follows: Science disciplines the mind and is an important element of a sound, basic education. If the student has reached eighth grade without receiving a foundation for the core content of high school, then there is an immediate educational problem that will indeed lead to later problems with international competitiveness. Without an objective test of knowledge, there can be no diagnosis, and without diagnosis there can be no rescue of the student. *Quod erat demonstratum.*

Not all tests are good tests (see George K. Cunningham's chapter on the Kentucky system). The opening example of a good question, taken from the NAEP Science Test, is, unfortunately, a rare exception. Consider the following four items from the same eighth-grade NAEP Science Test. These are numbers eight

through eleven out of a group of thirteen questions that refer to a diagram of a pond ecosystem:[2]

8. If all of the small fish in the pond system died one year from a disease that killed only the small fish, what would happen to the algae in the pond? Explain why you think so.

What would happen to the large fish? Explain why you think so.

9. Suppose that one spring a new type of large fish was put into the pond. So many were put in that there were twice as many fish as before. By the end of the summer, what would happen to the large fish that were already in the pond? Explain why you think these new large fish would have this effect.

10. If a rainstorm washed some fertilizer from a nearby field into the pond, what would happen to the algae in the pond system after one month? Why do you think the fertilizer would affect the algae this way?

11. What effect would the fertilizer have on the bacteria in the mud at the bottom of the pond after one month? Why do you think the fertilizer would affect the bacteria this way?

These questions do not have simple answers, nor do they even have single correct answers. The scoring guide is poorly constructed, giving the greatest reward to answers that are superficial. In question 8, the student is expected to write that the large fish will starve when the small fish die (and not be tempted to think that the fish might switch their diet to other animals pictured in the diagram, such as frogs and insects). In question 9, the student is expected to write that the "new type of large fish" that

is added to the pond will compete with the old type of large fish. This is not only an unwarranted assumption about resource usage, but doubly confusing because the large fish still ought to be dead from the previous question. When a rainstorm washes "some fertilizer" into the pond in question 10, the student is expected to assume that the concentration of nutrients for the algae increases and not that the accompanying rainwater causes an overall dilution of nutrients. And finally, in question 11 the student is to believe that the bacteria in the mud at the bottom of the pond will be greater in number one month after the rainstorm and not hesitate to make such a prediction with so little data at hand. A correct answer on question 10 is critical for answering question 11, and this lack of independence between questions is a serious problem with the way the NAEP Science Test was constructed.

Question 11 verges on being unanswerable, for scientist and student alike. Approximately 69 percent of U.S. students gave answers that were deemed "incorrect" or "unsatisfactory" by the scoring panel, 21 percent gave answers that were considered "partial," and 10 percent were scored as "omitted item" or "off task." Last but not least, the percentage of eighth-grade students in the nation with "complete" answers on question 11 was reported to be 0 percent (after rounding). This is a strong indication that the question is defective in some way and that it should have been excluded from the test form after scientific or psychometric review.

These kinds of testing defects may arise, in part, from the overpowering interest of educators in what they call higher-order thinking and conceptual understanding. A straightforward question on a test, especially one that has a single correct answer, is likely to be discounted as "mere recall." This attitude leads to the administration of test items that are fundamentally superficial, though they may hold the pretense of showing conceptual understanding.

For example, on the eighth-grade NAEP Mathematics Test, the following problem appears:

> 6. A poll is being taken at Baker Junior High School to determine whether to change the school mascot. Which of the following would be the best place to find a sample of students to interview that would be most representative of the entire student body?

A) An algebra class
B) The cafeteria
C) The guidance office
D) A French class
E) The faculty room

This particular question is classified by the NAEP writers as one that measures "students' conceptual understanding," though many readers might rightly wonder why it belongs on a mathematics test. Regarding this question, the NAEP writers go on to explain that:

> Students demonstrate conceptual understanding in mathematics when they provide evidence that they can recognize, label, and generate examples of concepts; use and interrelate models, diagrams, manipulatives, and varied representations of concepts; identify and apply principles; know and apply facts and definitions; compare, contrast, and integrate related concepts and principles; recognize, interpret, and apply the signs, symbols, and terms used to represent concepts.[3]

As to the aspects of data analysis, statistics, and probability being tested, they write, "This question also focuses on the subtopic of using measures of central tendency (that is, mean, median, range) to describe statistical relationships."[4] The test authors seem to have overestimated the mathematical value of the question and overlooked its apparent technical defects. The wording of the test item is likely to be misinterpreted by many students, because not all schools have each of the five named locations, nor would all use these exact titles for the sites (a cafeteria might be called a lunch room). Of those schools that even have a mascot, not all would be so progressive as to ask students for their opinions, which may explain why "The faculty room" was the most common incorrect answer nationally. Not all students are in settings in which the principal or teacher would even permit "An algebra class" or "A French class" to be interrupted by such a frivolous poll, and that may also have affected the students' thinking.

Just as there is grade inflation in the classroom, there seems to be cognitive inflation in test frameworks and in state and national

standards of learning. There is one peculiar characteristic of these types of writings, for example, the just-quoted passage from the NAEP writers, and that is the feverish use of action verbs. Students are to "recognize, label, and generate," or "compare, contrast, and integrate," or "recognize, interpret, and apply." The swarms of verbs usually arrive in groups of three or more, like horsemen of the apocalypse, and they drive away the teaching and testing of foundational knowledge. The recommendations are sometimes explicit on this matter, as in the National Science Education Standards, which call for "less emphasis on knowing scientific facts and information" and "more emphasis on understanding scientific concepts and developing abilities of inquiry."[5]

Ideally, test items would be grounded in a wide range of cognitive levels, but would also be consistent with the goals of scientists and mathematicians to explain real or abstract structures in the simplest terms. On the NAEP test items, half of the student's job may be untangling the awkward English. For example, in the aforementioned NAEP Science Test question regarding stocking a pond with a new type of large fish, the second sentence reads: "So many were put in that there were twice as many fish as before." In the NAEP Mathematics Test for eighth-grade students, the following question appears:[6]

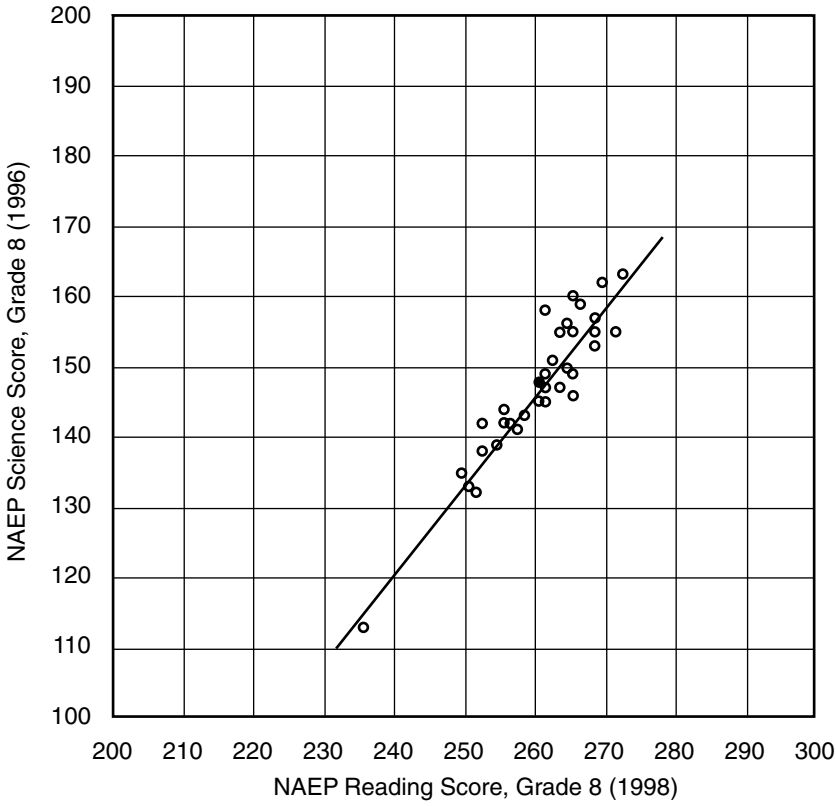
> There are 50 hamburgers to serve 38 children. If each child is to have at least one hamburger, at most how many of the children can have more than one?
>
> A) 6
> B) 12
> C) 26
> D) 38

This is cognitively a bit more difficult to solve than the question "50 - 38 = ?," but much of the burden for the student is in understanding the language rather than understanding the mathematics.

It may be fashionable to test students using word problems that are believed to measure higher-order thinking and to demand a constructed response rather than a selected response, but these practices may affect the validity of the test for students

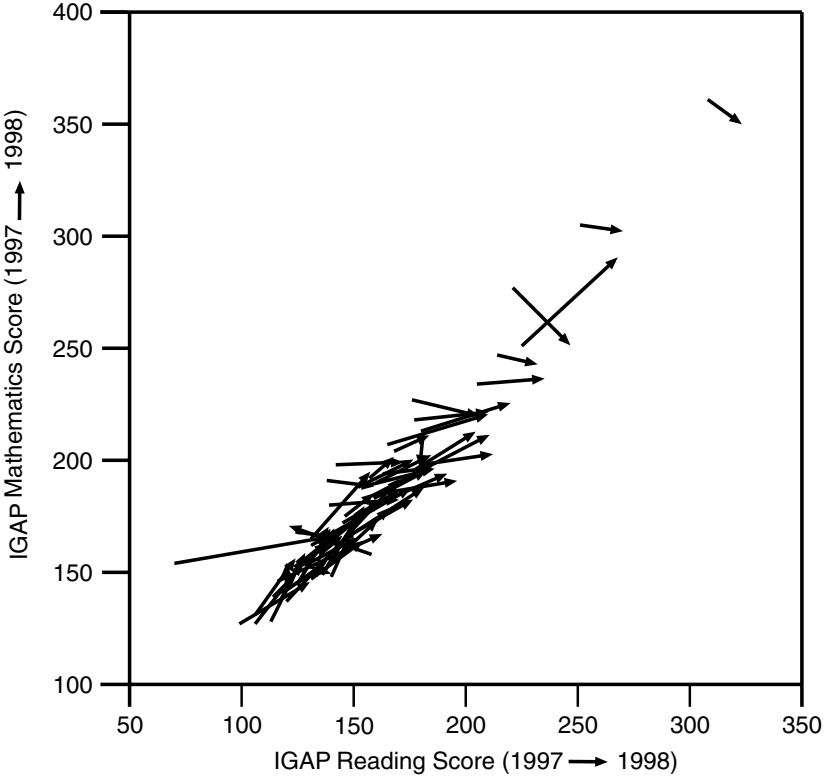
who have limited reading and writing skills. Validity, to a psychometrician, means that the test "actually measures the traits, knowledge, or skills it is intended to measure."[7] A test of science or mathematics that requires extensive reading or writing is partly a test of the intended content area and partly a test of the language skills needed to comprehend and complete the test form. The correlation shown in Figure 4.1 between state-level scores on the eighth-grade NAEP Science Test (1996) and the eighth-grade NAEP Reading Test (1998) is significant ($r = 0.93$, $n = 36$) and reproducible between test administrations.[8] A similar significant correlation exists between state-level performance on the eighth-grade NAEP Mathematics Test (2000) and NAEP Reading Test (1998) scores ($r = 0.94$, $n = 34$).

FIGURE 4.1 NAEP Reading and Science Scores Are Correlated



Other standardized tests yield similar findings. For example, in the Chicago Public High Schools the correlation between reading and math scores on the Illinois Goals Assessment Program (IGAP) test was significant in 1997 and 1998, with correlation coefficients of r = 0.95 and r = 0.97, respectively. In Figure 4.2, the change in reading and math scores for each school is shown in the form of individual arrows, with the tail of the arrow representing the reading and math scores in 1997 and the arrowhead representing the same paired data in 1998.[9] Of the sixty schools reporting data in both years, only one showed an increase in math score without an accompanying increase in reading score. It seems unlikely that pure mathematical ability is so closely tied to reading ability, and more probable that learning math (which may require

FIGURE 4.2 Correlated Improvement in Reading and Math IGAP Scores, Chicago Public High Schools, 1997–1998



reading the textbook) and taking an exam in math (which may require comprehension of the test) is sensitive to language skills.

A similar effect may be at play in the correlation between reading and science scores or all three may be under the control of an unknown general variable, some sort of common currency of academic ability.[10] It is unlikely to be that simple, and policy makers should carefully consider the more probable sources of the correlation. If the reading and writing demands of the test are more significant than the mathematics and science challenges, then these tests are partially invalid because they do not measure what they purport to measure. Alternatively, the correlation may show that reading ability is a strong determinant of the ability to learn other subjects. If students learn most of their science and math by reading, then constructivist curricula that rely heavily on hands-on activities and other manipulative projects may not be fully effective.

The awareness of constructivist methods of teaching is fairly widespread, as 84 percent of the eighth-grade math teachers reported being at least "somewhat knowledgeable" about the National Council of Teachers of Mathematics (NCTM) Curriculum and Evaluation standards in 1996, and 65 percent reported having attended some sort of professional development workshop or activity designed to help implement those standards.[11] Fourth-grade teachers were less familiar with the NCTM standards, with only 55 percent reporting that they are "somewhat knowledgeable" and only 40 percent reporting attending an NCTM-aligned workshop or activity[12] (though fourth-graders performed relatively better according to the international comparison done in the Third International Mathematics and Science Study [TIMSS]).

Among eighth-grade math teachers, 67 percent report having their students solve problems in groups at least once or twice a week, and 94 percent reported a frequency of at least once or twice a month.[13] One must be wary of drawing any firm conclusions from these types of survey data, as they may depend on perception as much as reality. For example, as shown in Table 4.1, when fourth-grade students and teachers are both asked how often the students discuss solutions to mathematics problems with other

students, the teachers report a relatively high frequency and the students report a relatively low frequency.

Perhaps these differ because the teachers report on the class average, whereas the student responses reflect the behaviors of individuals. Alternatively, as this question regards a teaching practice that is in high fashion, perhaps some teachers respond in the way they think they ought to and not in a way that reflects their own practice. One might find these data to be no more credible than the body weights people claim for their driver's licenses! If departments of motor vehicles want the truth, they must start weighing license applicants when their pictures are taken, and if educational policy makers want the truth, they must videotape the teaching (see the chapter by Alan Siegel on the TIMSS videotape studies).

Within an educational setting, it should be considered that some populations of students may be more sensitive to unsuitable teaching methods than others. The aforementioned correlation in reading and math ability in the Chicago Public Schools appears

TABLE 4.1

Question: **In this mathematics class, how often do you [the students] discuss solutions to mathematics problems with other students?**

	STUDENT RESPONSES[59]		TEACHER RESPONSES[60]	
Responses	*Prevalence*	*NAEP Score of Students by Response Type*	*Prevalence*	*NAEP Score of Teachers' Classes by Response Type*
Never or hardly ever	33%	222	6%	219
Once or twice a month	18%	227	22%	221
Once or twice a week	29%	224	37%	221
Almost every day	19%	217	35%	227

less strong in the schools performing in the top 10 percent. Four out of the top six schools showed a decrease in math score accompanied by an increase in reading score (these arrowheads point to the "southeast" on the graph, rather than to the usual "northeast"). Perhaps these students have reached a threshold reading ability at which comprehension of the textbook and examination is no longer problematic, so reading and math scores become independent variables. Alternatively, the top-performing students may have been negatively affected by the National Science Foundation's (NSF) Systemic Initiative Grant that was implemented at that time in the Chicago schools and that embraced the constructivist teaching methods recommended by the NCTM.[14]

According to the science teachers answering survey questions on the 1996 NAEP Science Test, hands-on methods of teaching science are widespread, with 86 percent of eighth-grade science teachers claiming to place "moderate" or "heavy" emphasis on developing students' laboratory skills and techniques.[15] Again, these responses are sensitive to perceptions and the psychology of surveys, but the reports from students raise some question about the value of hands-on science investigations in the classroom.

As Table 4.2 shows, student scores appear to decrease as the frequency of designing and carrying out science investigations increases, and it may be because these activities and projects do not involve much reading. Perhaps the best hands-on science program would be one in which students can get their "hands on" a good textbook.

TABLE 4.2

Question: When you study science in school, how often do you design and carry out your own science investigations?[61]

Responses	*Prevalence*	*Student NAEP Science Score*
Never or hardly ever	63%	151
Once or twice a month	23%	151
Once or twice a week	10%	142
Almost every day	5%	137

In California, poor reading skills have been the legacy of the whole-language movement, and while the picture is now becoming brighter for students in lower elementary grades, many of the students entering middle/junior high and high schools are deficient in reading–language arts skills. The numbers of students enrolled in Reading Improvement/Developmental Reading courses has increased dramatically in California, from 183,422 in the school year 1997–98 to 481,950 in 2002–3.[16] In seventh grade nearly one out of every four students is enrolled in this class. The effect of this problem on science instruction is difficult to measure because many districts continue to use hands-on kits for science instruction that have minimal reading materials for students. In 2002, California adopted the Science Curriculum Framework, which sets guidelines for K–8 instructional materials. With this document come new policies for teaching science.[17] The state-adopted science materials in grades K–8 are required to have expository text and cannot be based on more than 20 percent to 25 percent hands-on activities. Also, the new state-adopted language arts materials are required to address those K–3 science content standards that lend themselves to instruction during the language arts time period.[18] This is intended to protect that instructional time needed to develop language skills and is likely to be a wise investment over time.

This movement in California educational policy was partly in response to the 1994 NAEP Reading scores, which placed the state near the bottom of the national standings. One of the benefits of state, national, and international testing is that it can provide jurisdictions that are falling behind with a sense of reality and can have a positive effect on the curriculum. Enrollment in integrated math and integrated science courses[19] dropped significantly in California between 1998 and 2002,[20] and this may have been a rational response to the NAEP Science and Mathematics tests and the TIMSS report in 1998. Despite their good performance in the fourth-grade TIMSS Science Test,[21] U.S. students scored closer to the international average in the eighth-grade exam[22] and well below average in the terminal year of secondary school.[23] The poor performance of high school

seniors in the United States was accompanied by an unrealistic view of their own prowess, and perhaps educational policy makers similarly had inflated opinions. For example, students in each nation were asked to respond on a Likert Scale to the statement "I have usually done well in mathematics,"—as Table 4.3 shows, their actual national test scores did not reflect their level of self-confidence.

In a broader look at the responses, 75.9 percent of the U.S. students either agreed or strongly agreed with the statement that they had usually done well in mathematics, and only 24 percent either disagreed or strongly disagreed. The United States may have taken the prize for "most deluded nation" on this particular matter, since their scores were so low; however the source of this delusion is not entirely clear. It may be that U.S. teachers are heaping students with undeserved accolades and top marks or that the national standards against which U.S. students are judged are woefully low. If U.S. students think that mathematics is primarily about passing out hamburgers and finding the best place to poll students about a new school mascot, then they may be truly surprised by the expectations placed on students in other countries.

The TIMSS test and survey results were reported by the Washington, D.C.–based National Center for Education Statistics (NCES), a federal entity that is part of the U.S. Department of Education. Examples of their publications include the *Pursuing Excellence* series of reports.[24] NCES work

TABLE 4.3

Statement: I [the student] have usually done well in mathematics.

TIMSS Nations	*Students strongly agreeing with the statement*[62]	*Students' average mathematics general achievement score*
Netherlands	13.4%	560
Sweden	16.2%	552
United States	23.6%	461
Cyprus	20.8%	446

should be distinguished from the TIMSS reports published by education researchers at Michigan State University (MSU), a group that calls itself the "U.S. TIMSS National Research Center."[25] The MSU group is funded by the NSF and publishes highly interpretive reports on U.S. student achievement, some of which predate the TIMSS study.[26] Examples of their work include *A Splintered Vision,*[27] which argued that U.S. science and mathematics curricula show an intention to cover many more topics than other nations' and that these topics are presented in a fragmented and unfocused way in textbooks. In addition to the content differences between the NCES and MSU reports, there is a significant difference in how the two groups reached conclusions and how they brought results to the attention of the public. Whereas the NCES reported with some restraint, "Our analysis of TIMSS data does not suggest any single cause of this level of U.S. performance,"[28] the MSU group opined, "What is surprising is not the profoundly disappointing results but rather failing to realize how predictable those results were given what we already knew."[29] Adding to the confusion, the MSU group had hired the New York–based public relations firm of Hill and Nolton, and the group's press release from East Lansing was timed for the exact hour that the U.S. Commissioner of Education Statistics made his statement, at 11:00 a.m. EST on February 24, 1998.

The MSU group had considerable success in getting their message to stick in the public's minds, with their claim that the U.S. curriculum is "a mile wide and an inch deep."[30] Such a statement has broad appeal because scientists and mathematicians may agree with the "inch deep" part, finding contemporary teaching to be superficial in content, and outcomes-based educators agree with the "mile wide" part, finding the academic curriculum in need of severe pruning. There are many factors that indubitably influence the performance of students in science and mathematics, and it should be remembered that curriculum is but one. Without compelling evidence of a cause-and-effect relationship, TIMSS scores cannot be used to measure the relative values of national curricula, and this is a

critical error in the development of the MSU group's methodology. It was simply taken *a priori* that an analysis of curriculum would provide invaluable information for interpreting achievement results.[31]

The problem with the MSU group's TIMMS report is far more than mere data-dredging, however. Of particular concern is the apparent lack of detachment of the researchers from the selection of data used. In *A Splintered Vision,* the MSU group stated that "the TIMSS curriculum analysis was based primarily on state curriculum frameworks or guides and on supporting opinion by experts in mathematics and science education."[32] As a general rule, researchers cannot avoid having predilections or hopes about their results, and that is why it is critical that they blind themselves to the process of selecting the data. In the MSU study, one of the principal investigators was also the U.S. National Research coordinator. As such, he was responsible for designating the panel of curriculum experts that would develop the General Topic Trace Mapping data for the United States[33] and was involved in selecting the U.S. curricular documents to be analyzed.[34] To explain this another way, a researcher should never poll himself and become a data point in his own study.

The selection and sampling of documents (curricular guides and textbooks) in the MSU study was a key element in defining its data set. The MSU report indicates that their group of researchers "drew an appropriate random sample of state curriculum guides in 1992–93," but this comment has this footnote attached to it: "Selecting documents for the U.S. Curriculum Analysis presented considerable challenges given the nature of curriculum policy making and textbook markets in this country."[35] Random sampling and selection cannot both have happened. A further statement, "Documents were sampled rather than surveyed exhaustively,"[36] suggests that once a curriculum guide was chosen for analysis, there may have been additional, potentially uneven methods for selecting the portions of the text to be reviewed and coded. In the final analysis, twenty-two mathematics curriculum guides derived from thirteen different states were used to represent the United States in the study.[37]

A second type of data entering into the MSU group's study, the "supporting opinion by experts," may similarly have been selectively chosen. Most nations were represented by a single individual[38] and that individual had primary control of the data and opinion representing his or her country. The representatives were often the National Research coordinators; however, there was considerable variation as to whether the representatives were national education ministers or academics. With variation in mathematics and science expertise, these appointed individuals might have varied widely in their perceptions of their native countries and the content fields of mathematics and science.

The coding of the curriculum guides and textbooks was a process by which the intended curriculum was evaluated by reviewers, and code numbers expressing the mathematics content were assigned.[39] The taxonomic framework containing the code numbers is the organizing structure for all of the data collected on the curriculum guides and textbooks and is divided into ten major groups, as shown in Table 4.4 (the deeper branches of the taxonomic tree are not shown).

TABLE 4.4

CONTENT

1.1 Numbers

- 1.1.1 Whole numbers
- 1.1.2 Fractions and decimals
- 1.1.3 Integer, rational, and real numbers
- 1.1.4 Other numbers and number concepts
- 1.1.5 Estimation and number sense

1.2 Measurement

- 1.2.1 Units
- 1.2.2 Perimeter, area and volume
- 1.2.3 Estimation and errors

1.3 Geometry: Position, visualization, and shape
- 1.3.1 Two-dimensional geometry: Coordinate geometry
- 1.3.2 Two-dimensional geometry: Basics
- 1.3.3 Two-dimensional geometry: Polygons and circles
- 1.3.4 Three-dimensional geometry
- 1.3.5 Vectors

1.4 Geometry: Symmetry, congruence, and similarity
- 1.4.1 Transformations
- 1.4.2 Congruence and similarity
- 1.4.3 Constructions using straight-edge and compass

1.5 Proportionality
- 1.5.1 Proportionality concepts
- 1.5.2 Proportionality problems
- 1.5.3 Slope and trigonometry
- 1.5.4 Linear interpolation and extrapolation

1.6 Functions, relations, and equations
- 1.6.1 Patterns, relations, and functions
- 1.6.2 Equations and formulas

1.7 Data representation, probability, and statistics
- 1.7.1 Data representation and analysis
- 1.7.2 Uncertainty and probability

1.8 Elementary analysis
- 1.8.1 Infinite processes
- 1.8.2 Change

1.9 Validation and structure
- 1.9.1 Validation and justification
- 1.9.2 Structuring and abstracting

1.10 Other content
- 1.10.1 Informatics

These ten major groupings of content are unevenly distributed across the mathematics field, and considerable weight is given to measurement, estimation, data representation, and validation.

Beyond the primary groupings of content, the secondary and tertiary taxonomic levels are also unevenly allotted. Under category *1.1 Number,* the five subcategories are divided into a total of twenty additional codes (not shown). Fine distinctions are drawn at this tertiary level between the content areas of common fractions (*1.1.2.1*), decimal fractions (*1.1.2.2*), relationships between common and decimal fractions (*1.1.2.3*), percentages (*1.1.2.4*), properties of common and decimal fractions (*1.1.2.5*), and rational numbers (*1.1.3.2*). The major category *1.1 Number* is the only one that was divided to this third taxonomic level, and consequently it represents nearly half of the framework branches at their highest degree of specificity. Algebraic issues (*1.6 Functions, relations and equations*) merits only two subcategories overall, whereas proportionality (*1.5 Proportionality*) is specified by four subcategories, geometry (*1.3 Geometry: Position, visualization, and shape* and *1.4 Geometry: Symmetry, congruence, and similarity*) by eight, and number (*1.1 Number*) by twenty-five.

The MSU group did not force a one-to-one mapping between the curriculum material and the taxonomic scheme so as to take into account the "interrelatedness of content."[40] Individual "blocks" of content, small analytical segments of the books, were coded using the taxonomy, as shown here by one of their examples:[41]

Problem: The product of 0.23 and 6.57 is closest to:

a. 0.0015
b. 0.15
c. 1.5
d. 15.0
e. 150

Framework categories assigned:

1.1.2.2 Decimal fractions
1.1.5.3 Estimating computations

It is not clear why a reviewer would reject other potential classifications, such as *1.1.1.2 Operations* or *1.1.5.2 Rounding and significant figures,* or why a second- or first-generation taxonomic level might not be selected to code the task, such as *1.1.5 Estimation* and number sense or simply *1.1 Number.* This was a source of irregularity in their coded data because it led to differences in how countries evaluated the content of their own books. Reviewers in Korea, New Zealand, Switzerland, and the United States, to name a few,

regularly used the a general level of specification (*1.1 Number*) to code data from seventh- and eighth-grade textbooks, whereas reviewers in other nations, such as Hong Kong, Cyprus, Spain, and Germany, applied the taxonomic framework differently.

The MSU group considered each taxonomic framework category to be a different "topic," meaning that a teacher who assigned the estimation problem shown above would have been instructing students in two different topics simultaneously (decimal fractions and estimating computations). Of course, the student would perceive it as a single-topic question—it is only the scheme developed by the MSU group that makes it two topics. By the taxonomy, a textbook in algebra might be coded as containing only three topics from *1.6 Functions, relations, and equations,* whereas another book that included an introduction to topics from *1.1 Number* might be coded as containing twenty-five. The MSU group reported that "the U.S. mathematics and science textbooks analyzed included far more topics than was typical internationally at all grade levels analyzed,"[42] but this statement depends on both the uneven taxonomic framework and the irregular methods by which the framework was applied during coding.

U.S. curriculum policy was characterized by the MSU group as being unfocused and incoherent; however, their method of aggregating the curriculum guides from thirteen states to represent the nation may have been the factor that created a lack of focus. The researchers in the MSU group did not seem to be unaware of the problem, as they wrote:

> Countries without national curriculum guides, but with multiple subsystems and their corresponding guides, could, when aggregated to the country level, produce longer durations for some topics. As a result, the country-level durations should not be interpreted as specific to individual students but rather to the country as a whole.[43]

No U.S. students experience the superimposed curricula of thirteen different states, and using that model as a representation of their schooling leads to a research artifact.

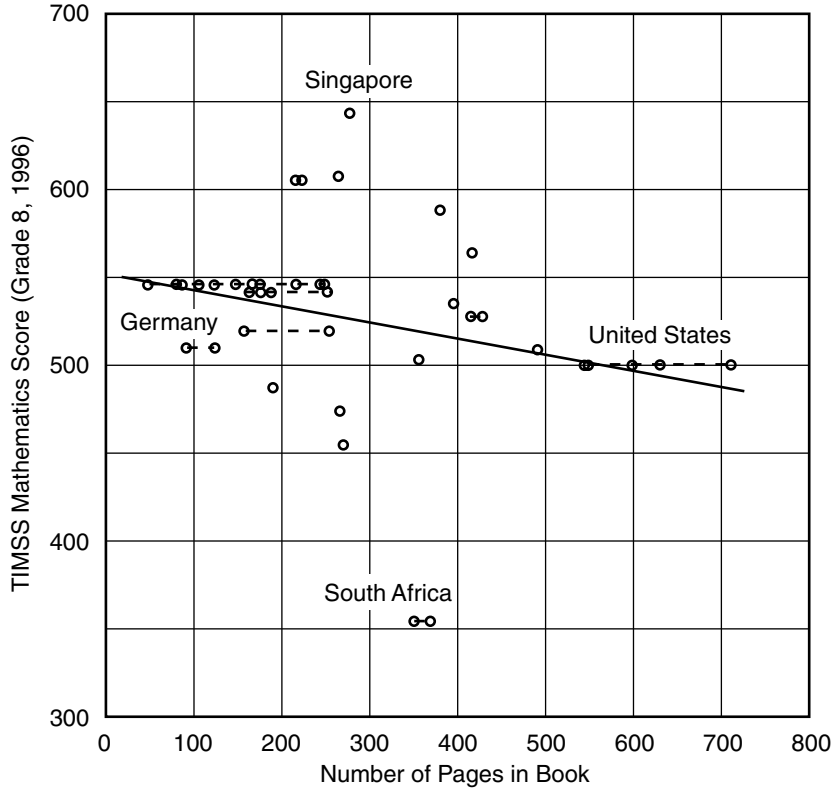
In remarks made at the Republican Governors' Conference in 1997[44] and to news organizations,[45] the MSU group voiced the opinion that a significant problem with U.S. textbooks is their sheer

length. It was said that because of the large books "teachers in the United States are forced to deal superficially with subjects and then review them again yearly, wasting valuable instructional time." It is a question worth reviewing, because textbook page counts, unlike topic counts, are an objective measure. The MSU group has made their data set public,[46] so the number of pages in math textbooks from different nations can be analyzed for potential correlation with student performance. When this is attempted with the eighth-grade math textbooks from nineteen nations, as shown in Figure 4.3, the data are widely scattered, and there does not appear to be any credible correlation ($r = -0.27$).[47] For example, the number of pages in the math book from Singapore (278 pages, TIMSS score 630) is not significantly different from the number of pages in the math books of South Africa (360 pages, TIMSS score 354), even though their scores are dramatically different. The United States (TIMSS score 500) gave students large books, ranging in size from 545 pages to more than 700 pages, but scored similarly to Germany (TIMSS score 509), which had books one-fifth that size.

It may be a false assumption that U.S. teachers begin with a class in September on the first page of a book and continue teaching until they get to the 700th page sometime in June. Many sections of the book may be skipped during teaching, and there may, in fact, be benefits to the unused pages. Commercial textbooks may contain extra units so that they can be better applied to the focused curriculum in each school and so that they might provide better support for teaching students of varied achievement levels. The sizes and weights of books may be something that every educational policy maker can love to hate, but until the U.S. Surgeon General finds that eighth-grade student spines have become deformed from the extra burdens in backpacks, it is reasonable to doubt that the extra pages are harmful.

Although little credibility ought be given to the MSU group's counting of textbook "topics" because their taxonomic framework was subjectively and irregularly applied during coding, it is remarkable that the data in their files do not even support their own conclusions. As shown in Figure 4.4, the relationship between TIMSS scores and topic counts is again a scattered one without credible correlation ($r = -0.21$), as it was for the textbook page counts from each

FIGURE 4.3 TIMSS Math Scores Do Not Correlate with Number of Pages in National Textbooks

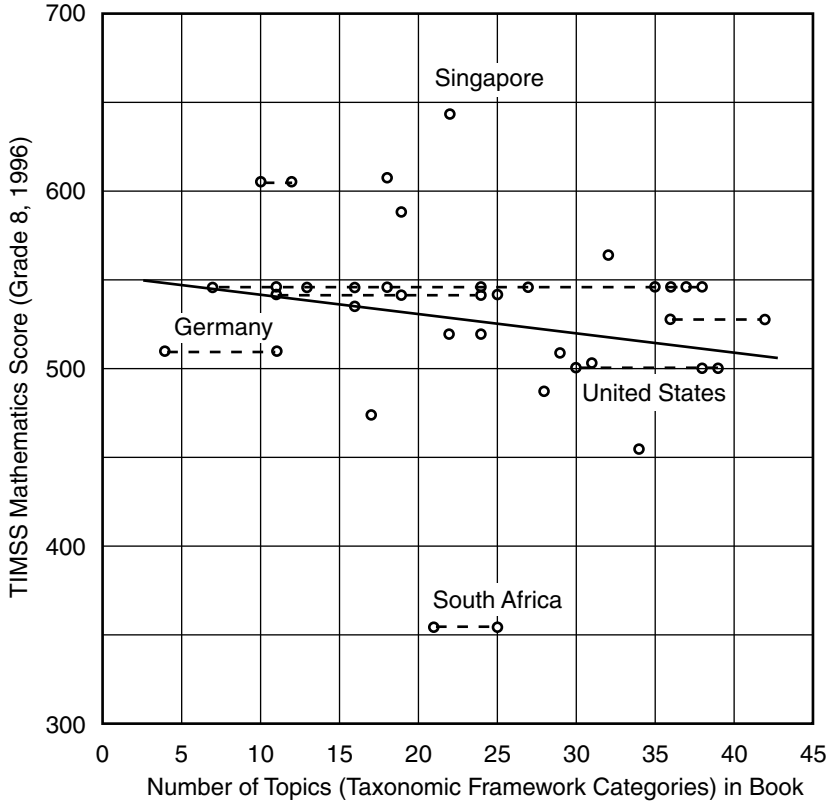


nation.[48] Singapore and South Africa, with widely different scores, have nearly identical numbers of topics in their textbooks, whereas the United States and Germany have nearly identical scores and widely different numbers of topics.

An attempt to correlate performance with coverage of specific topics would also fail. For example, Portugal (with a TIMSS score of 454) has considerably more textbook coverage of topics in the field of algebra than Singapore (with a TIMSS score of 630).[49]

The MSU group observed that according to its data, Japanese eighth-grade math books had long sequences of unbroken coverage of topics, whereas U.S. books tended to have shorter sequences on a single main topic and tended to break that sequence by attending to a different topic. A typical segment of twenty-five sequential blocks

FIGURE 4.4 TIMSS Math Scores Do Not Correlate with Number of Topics in National Textbooks



of data from a single U.S. and Japanese eighth-grade mathematics book are in Table 4.5 to illustrate this point. The U.S. reviewer encoded the twenty-five blocks from the U.S. book with one to two topics per block, and the sequence covers six different topics in total. The Japanese reviewer, on the other hand, gave each sequential block only one topic code per block, and within the twenty-five-block segment only one topic is covered.[50]

Both books are presenting the topic of algebra (equations and formulas). However, the reviewer of the U.S. book did not consistently encode the blocks as algebra (that is, the topic code *1.6.2 Equations and formulas* was omitted for blocks 5, 6, 11–18, 21,

TABLE 4.5 Twenty-five Sequential Topic Data Blocks from a U.S. Textbook and a Japanese Textbook

Block	*United States book*	*Japanese book*
1	1.3.1 (Coordinate geometry); 1.6.2 (Equations and formulas)	1.6.2 (Equations and formulas)
2	1.3.1 (Coordinate geometry); 1.6.2 (Equations and formulas)	1.6.2 (Equations and formulas)
3	1.1.3.3 (Real numbers); 1.6.2 (Equations and formulas)	1.6.2 (Equations and formulas)
4	1.1.3.3 (Real numbers); 1.6.2 (Equations and formulas)	1.6.2 (Equations and formulas)
5	1.1.3.3 (Real numbers)	1.6.2 (Equations and formulas)
6	1.1.3.3 (Real numbers); 1.3.1 (Coordinate geometry)	1.6.2 (Equations and formulas)
7	1.1.3.3 (Real numbers); 1.6.2 (Equations and formulas)	1.6.2 (Equations and formulas)
8	1.1.3.3 (Real numbers); 1.6.2 (Equations and formulas)	1.6.2 (Equations and formulas)
9	1.1.3.3 (Real numbers); 1.6.2 (Equations and formulas)	1.6.2 (Equations and formulas)
10	1.1.3.3 (Real numbers); 1.6.2 (Equations and formulas)	1.6.2 (Equations and formulas)
11	1.1.2.2 (Decimal fractions)	1.6.2 (Equations and formulas)
12	1.1.2.2 (Decimal fractions)	1.6.2 (Equations and formulas)
13	1.1.2.1 (Common fractions)	1.6.2 (Equations and formulas)
14	1.1.2.1 (Common fractions)	1.6.2 (Equations and formulas)
15	1.1.2.1 (Common fractions)	1.6.2 (Equations and formulas)
16	1.1.2.1 (Common fractions)	1.6.2 (Equations and formulas)
17	1.1.2.1 (Common fractions)	1.6.2 (Equations and formulas)
18	1.1 (Numbers)	1.6.2 (Equations and formulas)
19	1.6.2 (Equations and formulas)	1.6.2 (Equations and formulas)
20	1.6.2 (Equations and formulas)	1.6.2 (Equations and formulas)
21	1.1.3.3 (Real numbers)	1.6.2 (Equations and formulas)
22	1.1.3.3 (Real numbers); 1.3.1 (Coordinate geometry)	1.6.2 (Equations and formulas)
23	1.1.3.3 (Real numbers); 1.6.2 (Equations and formulas)	1.6.2 (Equations and formulas)
24	1.1.3.3 (Real numbers); 1.6.2 (Equations and formulas)	1.6.2 (Equations and formulas)
25	1.1.3.3 (Real numbers); 1.6.2 (Equations and formulas)	1.6.2 (Equations and formulas)

and 22). These differences in evaluation could be a result of different reviewing styles.

For example, if two individuals read Gertrude Stein's quotation "A rose is a rose is a rose" and evaluated it for its coverage of "roses," their conclusions might depend on their perspective. One reviewer might read the quotation one word at a time and think it fragmented, since "roses" were only intermittently addressed, and only three out of the eight words were on the topic. Another reviewer might read the phrase from a broader perspective and decide that all eight words were part of an ongoing discussion of roses. By the same token, an exposition on equations of functions might present a graph of a function on the real number plane, and one reviewer might think it part of the same topic (*1.6.2 Equations and formulas*), and another might code it as an interruption (*1.1.3.3 Real numbers* and *1.3.1 Coordinate geometry*). This appears to have been handled differently in each country and perhaps was related to the mathematics expertise of the reviewers. A reviewer without adequate expertise might read algebra books in much the same way as a "word-by-word automaton" would read Gertrude Stein.

In the evaluation of U.S. and Japanese eighth-grade textbooks, approximately 10 percent and 29 percent of the data blocks are coded as *1.6.2 Equations and formulas,* respectively.[51] The U.S. books' data blocks are a bit scattered (as in "a rose is a rose is a rose") compared with the Japanese books (as in "a a a rose rose rose is is"), but just as the complexion of the fairest Hollywood starlet might look coarse in an extreme close-up, the fragmentation of topics in these books needs to be studied from the right distance. Is the scattering of topics between data blocks real, or is it an artifact of noise created by the review? Data blocks typically represent only one-quarter to one-sixth of one textbook page in the MSU study, and this may be too fine-grained an analysis. A more appropriate view of the text might be to look at segments that amount to 1 percent of its length, which would amount to a few instructional days. If a topic such as *1.6.2 Equations and formulas* appears intermittently in the U.S. book, but frequently

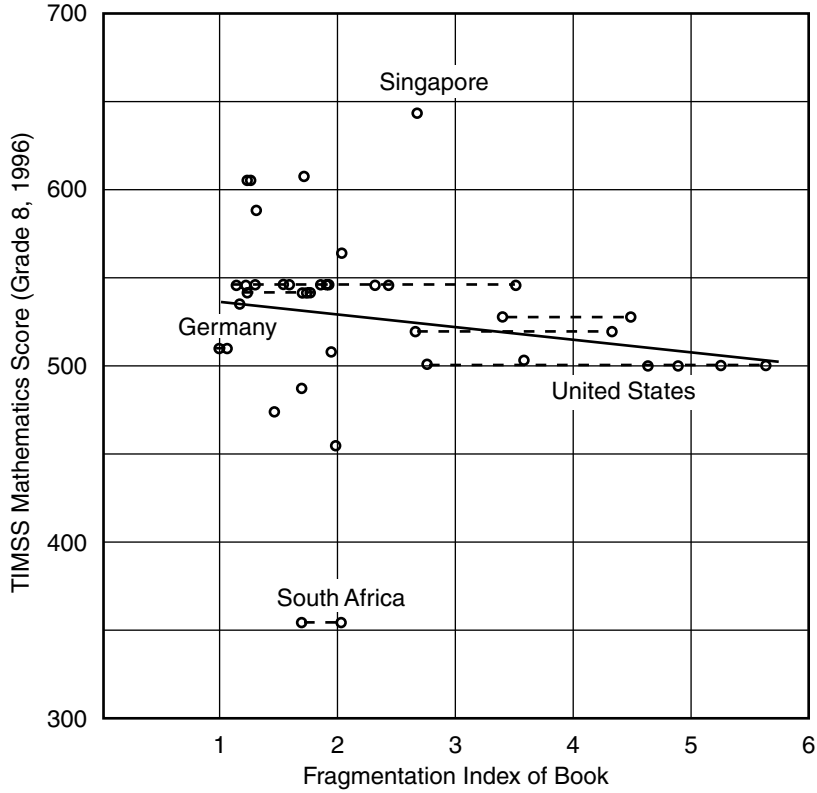
enough that a class is still effectively "on topic" for days, then the book may have a fairer complexion than was alleged.

In a reanalysis of the MSU group's data, 40 percent of the segments in the U.S. eighth-grade text contain the code *1.6.2 Equations and formulas,* with each "segment" taken to be 1 percent of the book.[52] That is to say, the "rose is a rose is a rose" looks considerably rosier at a reasonable distance than it does in the MSU group's close-up view of data blocks. Although it is possible that this indicates some sort of "complex signature"[53] of fragmentation in U.S. books, the chances of inconsistent treatment by different reviewers seems far greater. It hardly matters what the cause may be, however, because it is relatively easy to demonstrate that this type of fragmentation does not correlate with the TIMSS scores. The aforementioned U.S. book showed 10 percent coverage of *1.6.2 Equations and formulas* at the data block level and 40 percent coverage at the segmental level, so a "fragmentation index" of 4.00 is calculated for that book (that is, the index is simply the ratio, 40 percent divided by 10 percent). The Japanese book showed 29 percent coverage of the topic at the data block level and 31 percent coverage at the segmental level, so its fragmentation index is 1.07.

Figure 4.5 shows the fragmentation indices from the eighth-grade math books of nineteen countries, and it is apparent that they do not correlate to student performance on the TIMSS assessment ($r = -0.16$).[54] The South African books are less fragmented than the Singaporean books, though South Africa and Singapore place at the bottom and top of the TIMSS scoring respectively. The U.S. students score approximately as well on the TIMSS as German students, though the MSU TIMSS data indicate that their books are the most and least fragmented, respectively.

In their report *A Splintered Vision*, the MSU group created the impression, by anecdotal data taken only from the U.S. and Japanese books, that there might be a relationship between topic fragmentation and TIMSS performance.[55] Had they presented more than this small selection of their data on this particular question, the lack of correlation would have been immediately obvious.

FIGURE 4.5 TIMSS Math Scores Do Not Correlate with Topic Fragmentation in National Textbooks



Summary and Recommendations

Testing in mathematics and science is a critical element of standards-based reform, but it is not without its shortcomings in practice. Individual standardized tests such as the NAEP need to be redesigned with greater care so that they are correct in content and are valid measures of knowledge and skills in their respective content areas. This may mean turning away from some fashionable types of questions that make it difficult to measure content knowledge and skills in isolation of more general language skills. The panels that construct these tests need to be stocked with the best content experts that the nation can provide. A working formula might be that for every ten panel members writing a test,

five should have Ph.D. degrees in the subject (not in "education of the subject") and be active contributors to their fields. Four of the panel members should be full-time K–12 teachers in the subject area, with at least ten years' experience at the grade level being tested, and one of the panel members should be an iron-fisted psychometrician who knows that the test must be valid as well as reliable.

The survey data that accompany the NAEP and TIMSS tests should be analyzed for trends, but problems with the perception and psychology of the respondents may make it difficult to arrive at meaningful conclusions. There is no single cause of poor U.S. performance, and policy makers need to be cautious about any claim to the contrary. In particular, the findings of the Michigan State University TIMSS U.S. National Research Center are not credible. Their research design and methodology were flawed, and their own collected data do not support their published conclusions.

The MSU report *A Splintered Vision* expressed a view that national standards and reform were leading to progress and a wish that there could be greater local and state adherence to national reform guidelines.[56] The president of the National Academy of Sciences was similarly hopeful about the "pretty impressive" performance of fourth-grade students on the TIMSS and called for the nation not to step backward because of the poor twelfth-grade results. He said:

> Let me remind you that reform begins with the national education standards, and those standards must be implemented in the form of instructional materials, teaching methods, and assessments. Where such change has been made, for example, in elementary schools, we have preliminary indicators that education reform is working in those early grades.[57]

Was he right that the fourth-grade TIMSS results were a positive indication that national standards had benefited these fortunate students, and that eighth- and twelfth-grade students might soon be equally impressive?

The question was tested in 1999, when the United States participated in a follow-up TIMSS Repeat (TIMSS-R) study of

eighth-grade students. This was a second look at the cohort of students who had been "pretty impressive" in 1995 and had now had four additional years of U.S. schooling. Unfortunately, by their TIMSS scores in eighth grade, they had sunk to the same level as the previous eighth-grade students of 1995.[58] U.S. students may simply travel along a well-worn path, from "pretty impressive" to "pretty mediocre." Making a better path for future students will depend on the collection of valid testing data and educational research that is open-minded and credible.

Notes

1. The NAEP test is developed by the National Assessment Governing Board. This question from the eighth-grade NAEP Science Test, Block: 2000-8S11 No.:12.
2. Questions from eighth-grade NAEP Science Test, Block: 2000-8S9, Nos.: 8–11.
3. Content classification for item on eighth-grade NAEP Mathematics Test, Block: 1996-8M3 No.:6.
4. Ibid.
5. National Research Council, National Science Education Standards (Washington, D.C.: National Academy Press, 1996), 113.
6. Questions from eighth-grade NAEP Mathematics Test, Block: 1992-8M5 No.:5.
7. Educational Testing Service: http://www.ets.org/aboutets/statist.html.
8. A correlation between the NAEP Science Test (2000) and NAEP Reading Test (1998) scores is similarly significant ($r = 0.91$, $n = 30$). The correlation can only be made using states that participated in both types of assessments and between years in which the tests were administered. The filled circle in Figure 4.1 represents the national average scores and was not part of the correlation coefficient r.
9. Data extracted from school-by-school reports provided by the *Chicago Times* web site, http://chicagotribune.com/ws/front/0,1413,66,00.html. Sixty high schools in the Chicago public schools system reported math and reading data in both years and were included. An additional twenty were excluded because they failed to report complete data for one or both years; these were typically charter schools and transition schools.
10. If it is really that simple, then one reading test would suffice for students each year, and a set of exchange rates could be used to predict accurately the scores on the other tests.
11. C. A. Shaughnessy, J. E. Nelson, and N. A. Norris, *NAEP 1996 Mathematics Cross-State Data Compendium for the Grade 4 and Grade 8 Assessment* (Washington, D.C.:

National Center for Education Statistics, 1997), Table 6.24.

12. Ibid., Tables 6.1 and 6.2.
13. Ibid., Table 6.37.
14. The NSF director has claimed that the math score increases in Chicago during this time period were a result of NSF Systemic Initiative funding, a questionable attribution given the correlation with reading that is shown. Rita R. Colwell and Eamon M. Kelly, *Science* 286 (1999):237.
15. National Center for Education Statistics, *NAEP 1996 Cross-State Data Compendium for the Eighth Grade Assessment (NCES98-482).* (Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement), Table 5.1.
16. A course designed to provide instruction in basic language skills, integrating reading, writing, speaking, and listening while placing greater emphasis on individual student progress. In the California Basic Educational Data System (CBEDS), this course is CBEDS code number 2100. Data on specific course enrollment provided by the California Department of Education: http://data1.cde.ca.gov/dataquest.
17. California Department of Education *Science Framework for California Public Schools* (adopted by the State Board of Education, February 2002), http://www.cde.ca.gov/
cfir/science/.
18. California Department of Education *2002 K–8 Reading Language Arts English Language Development Adoption Criteria* (adopted by the State Board of Education, December 1999), http://www.cde.ca.gov/cfir/rla/2002criteria.pdf.
19. Integrated Mathematics 1 (CBEDS code 2425) course description: The course content includes functions, algebra, geometry, statistics, probability, discrete mathematics, measurement, number, logic, and language. The course emphasizes mathematical reasoning, problem solving, and communication through integration of the various strands, connections with other subject areas and real-life applications, use of technology, and exploratory and group activities. The course emphasizes algebra. Integrated Science 1 (CBEDS code 2626) course description: First-Year Coordinated/Integrated Science draws from the principles of several scientific disciplines—earth science, biology, chemistry, and physics—and organizes the material around thematic units. Common themes include systems, models, energy, patterns, change, and constancy. Students investigate applications of the theme using appropriate aspects from each discipline. http://www.cde.ca.gov/demographics/coord/curriculum/subject-table.htm.
20. 1998 course enrollment figures for ninth grade: Integrated Mathematics – 69,208 students; Integrated Science I – 96,858 students. 2001 course enrollment figures

for ninth grade: Integrated Mathematics I – 22,695; Integrated Science I – 73,353. Data extracted from California Department of Education, demographic data files. http://www.cde.ca.gov/demographics/files/cbedshome.htm.

21. U.S. Department of Education, National Center for Education Statistics, Pursuing Excellence: A Study of U.S. Fourth-Grade Mathematics and Science Achievement in International Context, NCES 97-255 (Washington, D.C.: U.S. Government Printing Office, 1997).
22. U.S. Department of Education, National Center for Education Statistics, *Pursuing Excellence: A Study of U.S. Eighth-Grade Mathematics and Science Teaching, Learning, Curriculum, and Achievement in International Context,* NCES 97-198 (Washington, D.C.: U.S. Government Printing Office, 1996).
23. U.S. Department of Education, National Center for Education Statistics, *Pursuing Excellence: A Study of U.S. Twelfth-Grade Mathematics and Science Achievement in International Context,* NCES 98-049 (Washington, D.C.: U.S. Government Printing Office, 1998).
24. Ibid.
25. Also variously called the National Research Center for the Third International Mathematics and Science Study, and the U.S. National Research Center (http://ustimss.msu.edu/).
26. For example, the Survey of Mathematics and Science Opportunities (SMSO) project. William H. Schmidt et al., *Characterizing Pedagogical Flow* (Boston: Kluwer Academic Publishers, 1996).
27. William H. Schmidt, Curtis C. McKnight, and Senta A. Raizen, *A Splintered Vision—An Investigation of U.S. Science and Mathematics Education* (Boston: Kluwer Academic Publishers, 1997).
28. Pascal D. Forgione Jr., Ph.D., U.S. Commissioner of Education Statistics, National Center for Education Statistics (NCES), press release, "On the Release of U.S. Report on Grade 12 Results from the Third International Mathematics and Science Study (TIMSS)" (February 24, 1998), http://nces.ed.gov/Pressrelease/timssrelease.html.
29. William H. Schmidt, U.S. TIMSS National Research Coordinator, press release, "Are There Surprises in the TIMSS Twelfth-Grade Results?" (Michigan State University: February 24, 1998), http://ustimss.msu.edu/12gradepr.htm.
30. Ibid.
31. Leonard J. Bianchi, Richard T. Houang, Jacqueline Babcock, and William H. Schmidt, *User Guide for the TIMSS International Curriculum Analysis Database* (East Lansing: TIMSS International Curriculum Analysis Center, Michigan State University, December 1998), 1–2; and Schmidt, McKnight, Raizen, *A*

Splintered Vision—An Investigation of U.S. Science and Mathematics Education, 11.

32. Schmidt, McKnight, Raizen, *A Splintered Vision—An Investigation of U.S. Science and Mathematics Education,* 13.
33. Bianchi, Houang, Babcock, and Schmidt, *User Guide for the TIMSS International Curriculum Analysis Database,* 4–5.
34. Ibid., pp. 1–8.
35. Schmidt, McKnight, Raizen, *A Splintered Vision—An Investigation of U.S. Science and Mathematics Education,* 13.
36. Ibid.
37. Ibid., Appendix B; and William H. Schmidt et al., *Many Visions, Many Aims* Volume I (Boston: Kluwer Academic Publishers, 1997), Appendix G.
38. Bianchi, Houang, Babcock, and Schmidt, *User Guide for the TIMSS International Curriculum Analysis Database,* 1–11.
39. Schmidt, McKnight, and Raizen, *A Splintered Vision—An Investigation of U.S. Science and Mathematics Education,* 1 (footnotes).
40. David F. Robitaille, *Curriculum Frameworks for Mathematics and Science,* (Vancouver: Pacific Educational Press, 1993), 44.
41. Ibid., 56.
42. Schmidt, McKnight, and Raizen, *A Splintered Vision—An Investigation of U.S. Science and Mathematics Education,* 5.
43. Schmidt et al., *Many Visions, Many Aims* Volume I, 71.
44. Remarks to the Republican Governors' Conference (Miami: November 21, 1997), http://timss.msu.edu/republican.html.
45. Debra Viadero, "Surprise! Analyses Link Curriculum, TIMSS Test Scores," *Education Week* (April 2, 1997): http://www.edweek.org/ew/ew_printstory.cfm?slug=27timss.h16.
46. Available in .zip compression format from http://timss.msu.edu/cdata.html.
47. Eighth-grade textbooks from nineteen nations representative of the full range of TIMSS scores were included in this analysis, with some nations submitting multiple textbooks that are connected in the diagram by dashed lines. The nineteen nations were Canada, Cyprus, Czech Republic, Germany, Hong Kong, Japan, Korea, the Netherlands, Norway, New Zealand, Portugal, the Russian Federation, South Africa, Singapore, Slovenia, Spain, Sweden, Switzerland, and the United States of America. The solid line represents linear regression analysis, without significance in correlation ($r = -0.27$).
48. Nineteen representative nations, as described in the previous note. Data points from nations submitting multiple textbooks are connected in the diagram by dashed lines. The solid line represents linear regression analysis, without significance in correlation ($r = -0.21$).

49. In Portugal 50 percent, and in Singapore 30 percent of eighth-grade book topics are subcategories of *1.6 Functions and Equations.*
50. The twenty-five sequential blocks from each country represent: United States, a twenty-five-block portion of lessons 24–26 from book 840-10088; Japan, a twenty-five-block portion of lessons 10–14 from book 392-0M182C.
51. The percentages refer to U.S. book 840-10081 and Japanese book 392-0M182C, although the comparison is general to other examples.
52. The statistical analysis was programmed and performed as follows: A "moving window" of width 1 percent of the sequential blocks per textbook was passed over the MSU TIMSS block data, with starting position incremented by one block for each window movement. The topic codes contained in each window and numbers of window positions containing each topic code were recorded for analysis. The ratio between the fraction of windows (book segments) containing each topic code and the fraction of blocks containing each topic code represents a fragmentation index, normalized so that an unfragmented topic in a book would yield a score of 1.00.
53. An expression used to describe national differences in fragmentation: Schmidt, McKnight, and Raizen, *A Splintered Vision—An Investigation of U.S. Science and Mathematics Education,* 130.
54. Nineteen representative nations, as previously described. Method of statistical analysis as previously described. Data points from nations submitting multiple textbooks are connected in the diagram by dashed lines. The solid line represents linear regression analysis, without significance in correlation ($r = -0.16$).
55. Schmidt, McKnight, and Raizen, *A Splintered Vision—An Investigation of U.S. Science and Mathematics Education,* 98–102. The data on fragmentation that dealt solely with Japanese and U.S. books is in Exhibits 37–38, pp. 99–101.
56. Ibid., 10.
57. Bruce Alberts, president, National Academy of Sciences, press release, "Twelfth-Grade Results from the Third International Math and Science Study (TIMSS)" (Washington, D.C.: National Press Building, February 24, 1998). The full context of the quoted portion: "This ['12th-Grade Results from the Third International Math and Science Study'] may appear to give credence to the position that education reform is not working. Quite the contrary is true. Let me remind you that reform begins with the national education standards, and those standards must be implemented in the form of instructional materials, teaching methods, and assessments. Where such change has been made, for example, in elementary schools, we have preliminary indicators that education reform is working in those early grades;

recall that the fourth-grade results from these international tests for U.S. students were pretty impressive. In fact, today's test results underscore the importance of education reform. We cannot use these results as justification for standing still or taking a step backward—that would be a great disservice to our children, who are looking to us to provide them with the skills needed for success."

58. National Center for Educational Statistics, *Third International Mathematics and Science Study—Repeat* (Washington, D.C.: U.S. Department of Education, 2001), http://nces.ed.gov/timss/timss-r/.
59. National Center for Education Statistics, *NAEP 1996 Cross-State Data Compendium for the Grade 4 and Grade 8 Assessment* (NCES98-481) (Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement, 1998), Table 6.11.
60. Ibid., Table 6.10.
61. National Center for Education Statistics, *NAEP 1996 Cross-State Data Compendium for the Grade 8 Assessment* (NCES98-482) (Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement, 1998), Table 5.9.
62. Third International Mathematics and Science Study, *Student Background Variables—Students in the Final Year of Secondary School* (INTMSL4=1) (Boston: International Study Center, Lynch School of Education, Boston College), Question CSBMGOOD, Location SQ3-22A. http://isc.bc.edu/timss1/database/pop3/POP3ALMN.ZIP.

Chapter 5

Telling Lessons from the TIMSS Videotape

Remarkable Teaching Practices As Recorded from Eighth-Grade Mathematics Classes in Japan, Germany, and the United States

Alan R. Siegel

Why Another Study?

The outstanding performance of Japanese students on the Third International Mathematics and Science Study (TIMSS) examinations, along with the accompanying TIMSS videotape classroom studies, have generated widespread interest in Japanese teaching practices. Unfortunately, despite this excitement, the

It is a pleasure to thank Clilly Castiglia and Kevin Feeley of the NYU Center for Advanced Technology and the Media Research Lab, who graciously provided the VHS frame processing. The author would also like to thank Professor Michiko Kosaka for resolving several questions about the actual Japanese as recorded in the excerpts.

majority of ensuing education analyses and policy reports seem to be based on incomplete portrayals of the teaching as documented on videotape. Part of the problem is that the teaching is remarkably rich. As a consequence, short summaries and even quotes from original sources sometimes fail to provide a balanced characterization of the actual lessons and can even be just plain wrong.

These are strong words, and especially so if they happen to allege serious errors and misunderstandings in widely cited and highly respected studies. However, these studies, despite being based on common sources of information, do sometimes contradict each other, so some of the assertions cannot be right. On the other hand, it is only fair to point out that there are just a few such contradictions; most of the conclusions are consistent across all of the studies. But we also concur with the overall theme: The lessons as recorded in Japan are masterful. The main—and crucial—difference is in understanding the kind of teaching that made these lessons so remarkable.

For example, it is widely acknowledged that Japanese lessons often use very challenging problems as motivational focal points for the content being taught.[1] According to the recent Glenn Commission Report:

> In Japan, . . . closely supervised, collaborative work among students is the norm. Teachers begin by presenting students with a mathematics problem employing principles they have not yet learned. They then work alone or in small groups to devise a solution. After a few minutes, students are called on to present their answers; the whole class works through the problems and solutions, uncovering the related mathematical concepts and reasoning.[2]

This chapter resolves the crucial classroom question that the other reports left unanswered: How in the world can Japanese eighth-graders, with just a few minutes of thought, solve difficult problems employing principles they have not yet learned?

Background

The Third International Mathematics and Science Study comprises an enormously complex and comprehensive effort to assess primary and secondary school mathematics and science educa-

tion worldwide. The examination phase began in 1995 with the testing of more than 500,000 students in forty-one countries[3] and continued with repeat testing (TIMSS-R) in 1999,[4] additional projects, and data analyses that are still a matter of ongoing research. As part of the TIMSS project, 231 eighth-grade mathematics lessons in Germany, Japan, and the United States were recorded on videotape during 1994–95. An analysis of these tapes, which includes a variety of statistics, findings, and assessments, was reported in the highly influential TIMSS Videotape Classroom Study by James Stigler and colleagues.[5] This study also provides a detailed description of its data acquisition and analysis methodologies. Subsequently, James Stigler and James Hiebert published additional findings in *The Teaching Gap,* which emphasizes the cultural aspects of teaching and offers suggestions about how to improve teaching in the United States.[6]

In addition, the project produced a publicly available videotape containing excerpts from representative lessons in geometry and in algebra for each of the three countries, along with a discussion of preliminary findings narrated by Dr. Stigler.[7] The excerpts of German and American lessons were produced in addition to the original 231 lessons, which are not in the public domain because of confidentiality agreements. For the Japanese lessons, disclosure permissions were obtained after the fact. The TIMSS videotape kit also includes a preliminary analysis of the taped lessons[8] that follows the procedures used in the actual study. In addition, the TIMSS project produced a CD-ROM with the same classroom excerpts.[9]

What the Video Excerpts Show

The video excerpts, it turns out, provide indispensable insights that complement the more widely cited studies. They are the primary source for the following analysis, which compares the assessments and conclusions of the many studies with the actual classroom events as documented on tape.

Geometry

The tape shows the Japanese geometry lesson beginning with the teacher asking what was studied the previous day. After working to

extract a somewhat meaningful answer from the class, he himself gives a summary: Any two triangles with a common base (such as *AB* in Figure 5.1) and with opposing vertices that lie on a line parallel to the base (such as the line through *C*, *D*, and *P*) have the same area because the lengths of their bases are equal, and[10] their altitudes are equal.

The teacher states this principle and uses his computer graphics system to demonstrate its potential application by moving vertex *P* along the line *CD*. The demonstration shows how to deform triangle *ABP* in a way that preserves its area. Next, he explains that this principle or method is to be the *foundation*[11] for the forthcoming problem, which he then presents. It is the following:

> Eda and Azusa each own a piece of land that lies between the same pair of lines. Their common boundary is formed by a bent line segment as shown. The problem is to change the bent line into a straight line segment that still divides the region into two pieces, each with the same area as before.

FIGURE 5.1 (letters *A* and *B* enhanced)

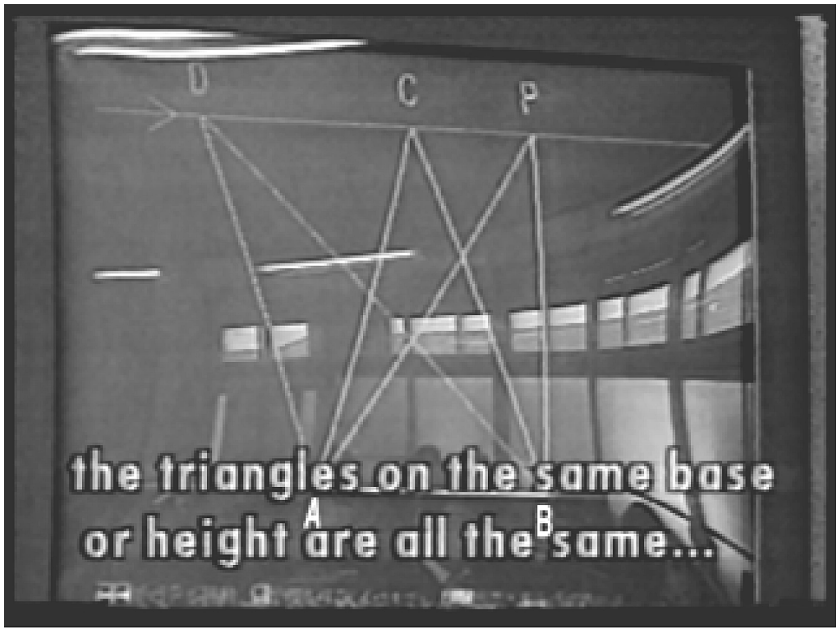
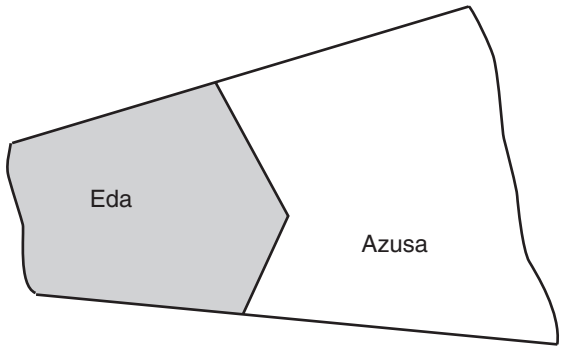


FIGURE 5.2



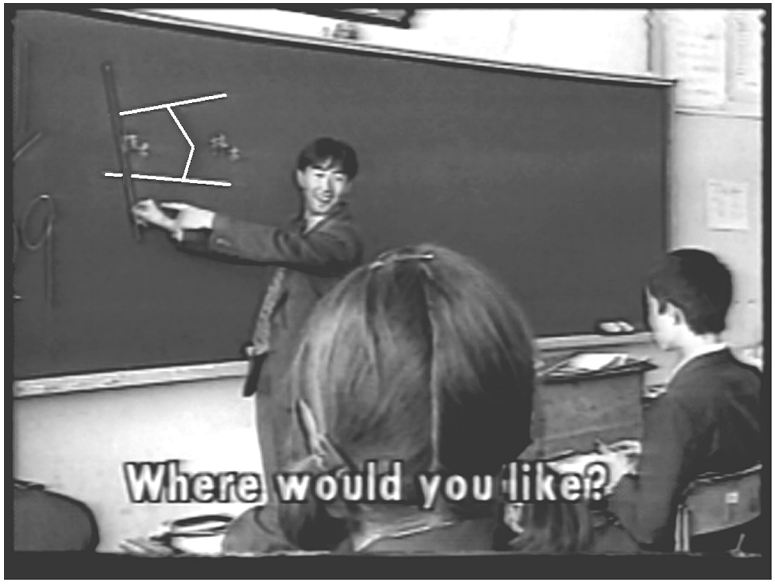
Despite the previous review, the problem is still going to be a challenge for eighth-graders, and it is fair to infer that the teacher understands this. In geometry, one of the most difficult challenges in a construction or proof is determining where to put the auxiliary lines. These lines are needed to construct the angles, parallel lines, triangle(s), and so on that must be present before a geometry theorem or principle can be applied to solve the problem. For the exercise in Figure 5.2, the key step is to draw two crucial auxiliary lines. One defines the base of a triangle that must be transformed in a way that preserves its area. The other is parallel to this base and runs through its opposing vertex.

So what should a master instructor do? The answer is on the tape.

After explaining the problem, the teacher asks the students to estimate where the solution line should go and playfully places his pointer in various positions that begin in obviously incorrect locations and progress toward more plausible replacements for the bent line. Now here is the point. With the exception of two positionings over a duration of about one second (which come shortly after the frame shown in Figure 5.3), none of his trial placements approximate either of the two answers that are the only solutions any student will find.

Rather, they are all suggestive of the orientation for the auxiliary lines that must be drawn before the basic method can be

FIGURE 5.3



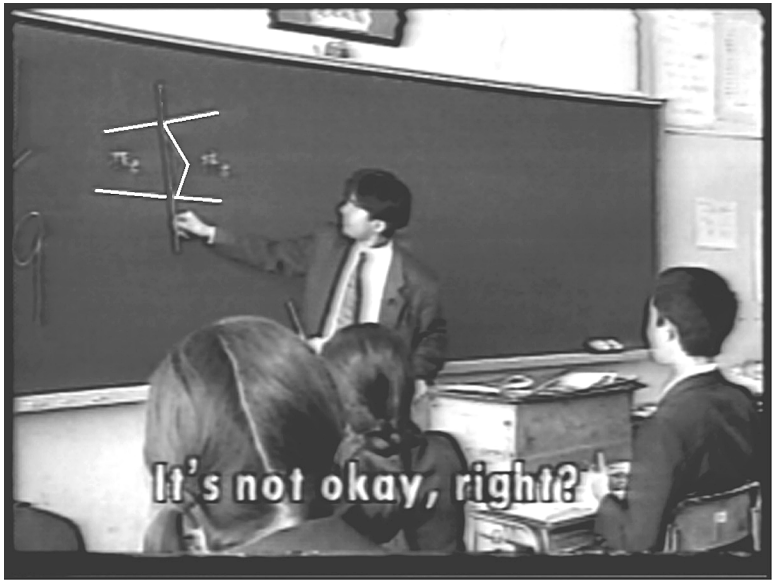
applied. He is giving subtle hints and calling the students' attention to the very geometric features that must be noticed before the problem can be solved. It is surely no accident that the teacher reaches two particular pointer placements more often than any other. One is shown in Figure 5.4. The other is parallel to this placement, but located at the vertex that forms the bend in the boundary between Eda and Azusa.

Only after this telling warm-up—the heads-up review of the solution technique necessary to get the answer and the seemingly casual discussion loaded with visual cues about what must be done—are the children allowed to tackle the problem.

But this is not the end of the lesson, and the students only get an announced and enforced three minutes to work individually in search of a solution.

As the children work, the teacher circulates among the students to provide hints, which are mostly in the form of leading questions, such as: "Would you make this the base? [The question is] that somewhere there are parallel lines, OK?"[12]

FIGURE 5.4

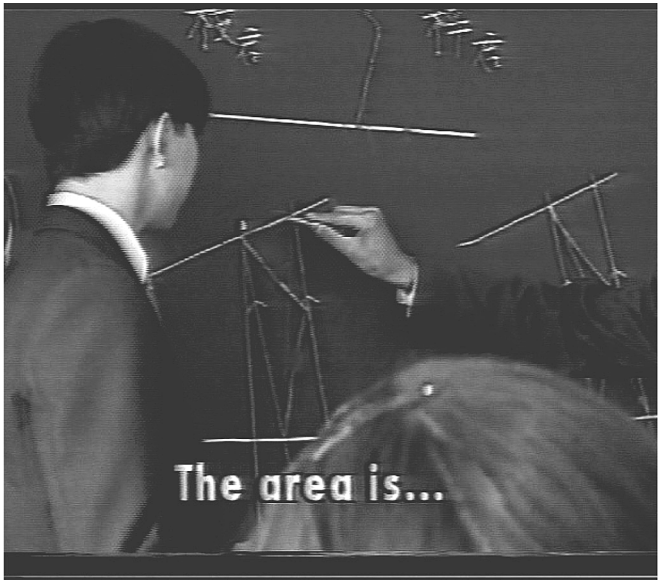


He then allocates an additional three minutes during which those who have figured out the solution discuss it with the other teacher. Weaker students are allowed to work in groups or use previously prepared hint cards. The tape does not show what happens next. The TIMSS documentation reports that students prepare explanations on the board (nine minutes).[13]

Then a student presents his solution. The construction is clearly correct, and he starts out with a correct explanation. However, when the time comes to find the solution, he gets lost and cannot see how to apply the area-preserving transformation that solves the problem. The teacher then tells him to use "the red triangle" as the target destination.

The advice turns out to be insufficient, and the teacher *steps in* (as shown in Figure 5.5) to redraw the triangle that solves the problem and calls the student's attention to it with the words "over here, over here." The student seems to understand and begins the explanation afresh. But he soon winds up saying, "Well, I don't know what I am saying, but . . . ." He then regains

FIGURE 5.5



his confidence, and the presentation comes to an end. A number of students say that they do not understand. Then another student explains her answer, but the presentation is omitted from the tape. According to the Moderator's Guide,[14] these two student presentations take less than three minutes altogether.

FIGURE 5.6

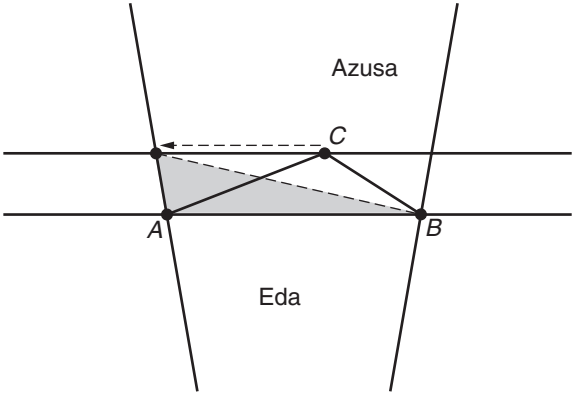
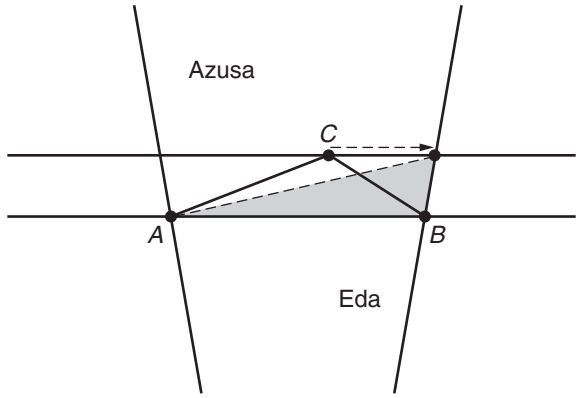


FIGURE 5.7



Next, the teacher explains how to solve the problem. There are two equivalent answers that correspond to moving vertex *C,* in the context of Figure 5.1, to the left or to the right. Both directions solve the problem, and he shows this. Such a duality should not be surprising, because the word problem is not described in a way that, in the context of Figures 5.1, 5.6, and 5.7, can distinguish left from right. For completeness, we show the two ways that the triangle transformation technique can be used to solve the problem. In order to make the connection between the review material and the follow-up Eda-Azusa exercise absolutely clear, the solution with its two versions have been rotated to present the same perspective as in Figure 5.1, which introduced this triangle transformation technique.

No one devised an alternative solution method.

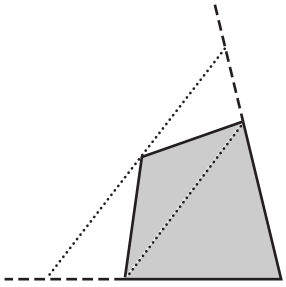
The lesson continues with the teacher posing a new problem that can be solved with the same technique. This time the figure is a quadrilateral, and the exercise is to transform it into a triangle with the same area. At this point, the basic solution method should be evident because the previous problem, as the teacher pointed out, also concerned the elimination or straightening of a corner in an area-preserving way.[15] However, added difficulty comes from the need to recognize that two consecutive sides of the quadrilateral should be viewed as representing the bent line of

FIGURE 5.8

Figure 5.2. Notice, by the way, that if each of the other two neighboring sides is extended as an auxiliary line, then the resulting figure is changed into a version of the Eda-Azusa problem. (See Figure 5.9.) Evidently, this exercise is well chosen.

The basic line-straightening method can be applied so that any one of the four vertices can serve as the point where the line bends, and this designated vertex can be shifted in either of two

FIGURE 5.9



directions to merge one of its two connecting sides with one of the auxiliary lines. The students again work individually for three minutes, then are allowed to work in groups, use hint cards, or ask the teacher.

The TIMSS documentation indicates that this joint phase lasts for twenty minutes and includes students presenting their answers. There are apparently eight such presentations, which were selected to illustrate all eight ways the basic method can be applied: There are four vertices that each can be moved two ways. Then the teacher analyzes these eight ways in greater depth and explains how they all use the same idea. All students remain seated during this portion of the lesson, and the teacher controls the discussion carefully and does almost all of the speaking.

An Analysis of the Teaching and Its Content

This lesson is nothing less than a masterpiece of teaching, and the management of classroom time was remarkable. Although many students did not solve the first problem of the day, the assignment certainly succeeded in engaging everyone's attention. The second problem was no giveaway, but it afforded students the chance to walk in the teacher's footsteps by applying the same ideas to turn a quadrilateral into a triangle. The teacher-led study of all possible solutions masked direct instruction and repetitive practice in an interesting and enlightening problem space.

Evidently, no student ever developed a new mathematical method or principle that differed from the technique introduced at the beginning of the lesson. In all, the teacher showed ten times how to apply the method. The lesson is an excellent example of how to teach problem solving because each successive problem required an ever deeper understanding of the basic proof technique. For homework, the teacher asked the students to transform a five-sided polygon into a triangle with the same area.[16]

Notice that this lovely problem variation hints at the use of induction: The way to solve it is to transform a five-sided figure into a quadrilateral, which can then be transformed into a triangle. The basic corner elimination scheme can now be seen to

work for any (convex) polygon so that any such n-sided polygon can be transformed into one with $n - 1$ sides and the same area, for $n > 3$.

It is also worth pointing out that the solution technique, which is a specific application of measure-preserving transformations, has additional uses. It appears, for example, in Euclid's proof of the Pythagorean Theorem (cf. Book I Prop. 47 of Euclid's *Elements*).[17] More advanced exercises of this type appear on national middle school mathematics competitions in China and regional high school entrance examinations in Japan. And it is not much of a stretch to suggest that measure-preserving transformations lie at the heart of those mysterious changes of variables in the study of integral calculus. All in all, the lesson is a wonderful example of the importance of a deep understanding of mathematics and its more difficult aspects.

Algebra

The Japanese algebra lesson begins with student-presented answers for each of the previous day's six homework problems.[18] These activities, along with the accompanying classroom discussion, are omitted from the excerpts.

Then the teacher presents a more challenging problem. It uses the same basic calculation method that the students have been studying, but needs one commonsense extension. The problem is this:

> There are two kinds of cake for sale. They must be bought in integer multiples; you cannot buy a fraction of a cake. The most delicious cake costs 230 yen, and a less tasty one is available for 200 yen. You wish to purchase ten cakes but only have 2,100 yen. The problem is to buy ten cakes and have as many of the expensive cakes as possible while spending no more than 2,100 yen.

It is clear that the students had already studied versions of the problem that would permit fractional units of cakes to be purchased. The reproduction of the six homework exercises as shown in the TIMSS Moderator's Guide confirms that the class was already experienced with the technical mechanics necessary to solve problems with inequalities.[19] It is also evident that they

had been studying word problems and the translation of word problems into equations and inequalities that can then be solved. Indeed, the teacher introduces the problem with the remarks, "Today will be the final part of the sentence problems."[20] Thus, it is fair to infer that the only difference between the cake problem and the material they had just reviewed is the requirement that the solution must comprise integer multiples of each cake.

After making sure that the students understand the problem, he asks them to devise a way to solve it. They get an announced and enforced three minutes.

Next, the teacher solicits solution approaches from the students. A student volunteers that she tried all possibilities. Her approach was to try ten cheap cakes, then nine cheap ones and one expensive one, and so on, until she had the best answer. However, she was unable to finish in the three minutes that the teacher allocated for the problem. The teacher emphasizes the point, and it will soon become clear that part of the lesson is to show that this unstructured approach is unsound.

He then briefly discusses another way to solve the problem. The approach, which is quite inventive, uses a notion of marginal cost. If we buy ten of the most expensive cakes, we exceed our budget by 200 yen. Trading in an expensive cake for a cheaper cake gives a net savings of 30 yen. Obviously, seven cakes have to be traded in, which shows that the answer is three expensive cakes and seven cheaper ones. As the teacher expected,[21] no student solved the problem this way.

Then he calls on another student, who explains how she set up the problem as an inequality, solved it as an equality, then rounded the number of expensive cakes down to the nearest lesser integer. As she explains the equation, he writes it on the board. Only a few students understand the explanation, and he asks for another explanation of the same process. In subsequent activities that are only summarized on the tape and in the Moderator's Guide, the teacher passes out a worksheet and works through a detailed analysis of the solution for the class.

After the detailed presentation, another problem of the same type was assigned, but with larger numbers. The teacher's words are telling:

> If you count one by one, you will be in an incredibly terrible situation. *In the same way that we just did the cake situation, set up an inequality equation by yourself* and find out . . . [the answer]. Because finding the answers one by one is hard, I wonder if you see the numerous good points of setting up inequality equations . . . .

The students work on the problem individually. After eleven minutes, the teacher went over the problem with the class. The video excerpts contain no group-based problem solving in this algebra lesson, and the Moderator's Guide confirms that none of the class time included problem solving in groups.

Each class ended with the teacher summarizing the solution technique that constituted the lesson of the day.

An Analysis of the Teaching and Its Content

Students never developed new solution methods. In the algebra class, the students were given the opportunity to learn firsthand why amorphous trial-and-error approaches (which seem to be encouraged by some of the latest reform programs) do not work. While the tape does not explicitly show how many students were able to solve the original cake problem in the allotted time, the student responses suggest that no more than four or five could possibly have succeeded. But the three minutes of struggle might well have served to make the lesson more purposeful.

From a mathematical perspective, the cake problem was designed to require a deep understanding of inequality problems and their solution. Mathematicians would say that when we solve a problem, we find all of the answers. If the cake problem had allowed fractional purchases and had simply required that altogether any mix of ten cakes be purchased for at most 2,100 yen, then the algebraic formulation would read

$$230x + 200(10 - x) \leq 2100$$

where x is the number of expensive cakes purchased and $10 - x$ is the number of the inexpensive ones. The problem would also require that x be nonnegative because you cannot buy negative quantities of cake. A little algebraic manipulation gives the solution as the interval

$$0 \leq x \leq \tfrac{10}{3}$$

Now, every x in this interval is a solution to the simplified problem, and every solution to the problem is in this interval. So if we want a special answer, the interval $[0, \frac{10}{3}]$ is the place to look. If we want the largest x, it is $\frac{10}{3}$. If we want the largest integer x, it is 3. And if we wanted the largest even integer, for example, we would look nowhere else than into $[0, \frac{10}{3}]$ to conclude that this answer is $x = 2$. Incidentally, a complete answer must also observe that the number of inexpensive items is nonnegative (which is to say that $x \leq 10$).

So this problem variant is more than a matter of common sense; it exposes students to a deep understanding of solutions to inequalities and the implications of real-world constraints. Moreover, the problem illustrates the idea of decomposing a complex exercise into a more basic problem whose solution can then be adapted to achieve the original objective.

In summary, the video excerpts feature challenge problems that cover fundamental principles, techniques, and methods of systematic thought that lie at the heart of mathematics and problem solving. As such, they ought to provide experiences that build a powerful foundation of intuition and understanding for more advanced material yet to come. As a derivative benefit, these problems are so rich they can be readily transformed into follow-up exercises for use as reinforcement problems in class and as homework.

Defining Terms: Discovery and Invented Methods

Many publications claim that the Japanese lessons teach students to invent solutions, develop methods, and discover new principles. For example, this view is expressed in the Glenn

Commission report[22] and is endorsed by the Videotape Study as well: "[In Japan, the] problem . . . comes first [and] . . . the student has . . . to invent his or her own solutions."[23] In fact, the Videotape Study reports that the fifty Japanese lessons averaged 1.7 student-presented alternative solution methods per class.[24] Yet the excerpts exhibit no signs of such activity. They contain just one student-devised solution alternative, and it failed to produce an answer.

These differences are fundamental, and they should be reconciled. Part of the difficulty is that students are unlikely to devise their own solutions when the time is limited, and the problems are so difficult that hints are needed. Moreover, the exercises seem to be designed to teach the value and use of specific techniques. Students would presumably have a better chance of finding alternative solution methods for less challenging exercises. And they would have an even better chance with problems that can be solved by a variety of methods that have already been taught. Examples might include geometry problems in which different basic theorems can be used and studies of auxiliary lines in which the exercises are designed so that different auxiliary lines build different structures that have already been studied. The Videotape Study illustrates alternative solution methods with the U.S. assignment to solve $x^2 + 43x - 43 = 0$ by completing the square and by applying the quadratic formula.[25] Of course, this problem directed students to use different methods they already knew. The example contains no hint of any discovery.

So the questions remain: Where are the alternative solution methods, and when do they demonstrate signs of student discovery?

The answers are in the Videotape Study. It presents actual examples that were used to train the data analysts who counted the "Student Generated Alternative Solution Methods" (SGSM1, SGSM2, . . .) in each lesson. These examples, it turns out, come from the geometry lesson in the video excerpts: The two student presentations for the Eda-Azusa problem are coded as SGSM1 and SGSM2.[26] Similarly, the second problem, in which each of four vertices could be moved in two directions, has the codings SGSM1-SGSM8. *Altogether, this lesson is*

counted as having ten student-generated alternative solution methods, even though it contains no student-discovered methods whatsoever. And the failed try-all-possibilities approach in algebra excerpts is counted as yet another student-discovered solution method.[27]

The Videotape Study also contains a partial explanation for the source of these judgments. It reports that the data coding and interpretation procedures were developed by four doctoral students—none of whom were in mathematics programs.[28] Moreover, the Videotape Study states that the project's supporting mathematicians only saw code-generated lesson tables and were denied access to the actual tapes.[29] It seems reasonable to infer, therefore, that they did not participate in the design of these coding practices. As for the question of invention, the Videotape Study explains: "When seatwork is followed by students sharing alternative solution methods, this generally indicates that students were to invent their own solutions to the problem."[30] There appears to have been a sequence of interpretations based on student presentations being very generously counted as student-generated alternative solution methods and, ultimately, as some kind of invented discoveries that might even depend on new principles the students had not yet learned.[31]

On the other hand, the contributions by the Japanese teachers received much less generous recognition. Yet in the defining examples of student discovery, the teachers—not the students—manage the ideas and lead the education process.

Additional Statistics from the TIMSS Projects

It is worth reiterating that in the Japanese lesson excerpts, each of the four exercises began with students working individually to solve the problem. Similarly, the Stigler-Hiebert analysis states, "Students rarely work in small groups to solve problems until they have worked first by themselves."[32] The detailed TIMSS Videotape Classroom Study contains no comparable statement and even implies otherwise: "[After the problem is posed, the Japanese] students are then asked to work on the problem . . . sometimes individually and sometimes in groups."[33] However,

not one of its eighty-six figures and bar charts documents instances where problems began with students working in groups. Chart 41 indicates that of the seatwork time spent on problem solving, 67.2 percent of the time comprised individual effort and 32.8 percent of the time was spent in group work.[34]

Another TIMSS study addressed this issue by collecting statistics for carefully balanced samples of eighth-graders. For each country, the sample base comprised approximately 4,000 students. Their teachers were queried about their classroom organization and whether most of the lessons had students working in small groups, individually, and/or as a class. Teachers also were asked if they assisted students in the classroom assignments. The results, which were weighted by the number of students in each responding teacher's class, are reproduced below (Figure 5.10) for the United States and Japan.[35]

The results show that Japanese lessons do not have significant numbers of small-group activities. In fact, American classes evidently contain four to six times as many such lessons. Of course, it should be noted that the data is based on questionnaires and depends, therefore, on the judgment of each respondent. The meaning of "most or every lesson" might have cultural biases, as might the definitions of "small groups" and "teacher assistance."

FIGURE 5.10

Country	Organizational Approach "Most of Every Lesson"				
	Work Together as a Class with Teacher Teaching the Whole Class	Work Individually with Assistance from Teacher	Work Individually without Assistance from Teacher	Work in Pairs or Small Groups with Assistance from Teacher	Work in Pairs or Small Groups without Assistance from Teacher
Japan	22	78	27	15	1
United States	r 22	r 49	r 50	r 19	r 12

An "r" indicates teacher response data available for 70–84% of students.

Still, these TIMSS statistics support the notion that the Japanese style of teaching is substantially different from many of the U.S. reform practices.

The Matter of Pedagogy

One such reform approach relies on discovery-based learning, which aims to have the students themselves discover mathematical principles and techniques. At first blush, the idea of discovery-based learning seems attractive. After all, we are more likely to recall what we discover for ourselves, and even if we forget such a fact, we should be able to rediscover it at a later date. According to Cobb and colleagues, "It is possible for students to construct for themselves the mathematical practices that, historically, took several thousand years to evolve."[36]

However, as with any idealized theory, the real issues are in the implementation practices.

- Judgments must determine how much classroom time should be allocated for students to discover the mathematics and must resolve the necessary tradeoffs among allocated time for guided discovery, for direct instruction, for reinforcement exercises, and for review.
- There must be detection and correction mechanisms for incorrect and incomplete "discoveries."
- There must be allowances for the fact that in even the best of circumstances, only a handful of students have any likelihood of discovering nontrivial mathematical principles.

The videotaped lessons from Japan show fundamental decisions that are startling and quite different from the reform practices in the United States. In the Japanese classes, the time allotted for the first round of grappling with problems was remarkably modest. Consequently, the remaining time was sufficient for student presentations to help identify conceptual weaknesses, for teacher-managed assistance and summations, and for follow-up problems designed to solidify understanding. However, because of the time

limitations and the difficulty of the problems, most students were learning via a model of "grappling and telling." That is, students would typically struggle with a tough problem in class, but not find a solution. They would then learn by being told how to solve it and would benefit from the opportunity to contrast unsuccessful approaches against methods that work.[37] There is no question that preliminarily grappling with a problem is both motivational and educational.[38] Similarly, discussion about why some approaches fail and why a solution might be incomplete, along with the exploration of alternative problem-solving techniques are all highly beneficial investments of time. But the use of grappling and telling creates yet another implementation issue, which is: Who should do the telling?

In some teaching practices, the theory of discovery-based learning is extended to include the notion of cooperative learning, which holds that the students should teach one another because they "understand" each other. However, both the TIMSS videotape and the data in Figure 5.10 show that the Japanese style of teaching is by no means purely or principally based on cooperative learning. Although students get a substantial amount of time to explain their solutions, the video excerpts show that Japanese teachers are by no means passive participants. Student explanations frequently need—and get—supervision, and students can be remarkably incoherent (cf. Figure 5.5) even when their solutions are absolutely perfect. When all is said and done, the teachers do the teaching—and the most important telling—but in an interactive style that is highly engaging and remarkably skillful.

According to Stigler and Hiebert, some lessons feature considerably more direct instruction or extended demonstrations while yet others demand that the students memorize basic facts.[39] Students might even be asked to memorize a mandate to think logically.[40] Evidently, the lessons do not follow a rigid pattern. If any theme is common to these approaches, perhaps it is that although the lessons vary depending on the nature of the mathematical content, they always engage the students in an effort to foster thinking and understanding.

Placing Japanese Teaching in the Context of U.S. Reform

The video excerpts show Japanese lessons with a far richer content than the corresponding offerings from the United States and Germany. According to the Videotape Study, the Japanese, German, and U.S. eighth-grade classes covered material at the respective grade levels 9.1, 8.7, and 7.4 by international standards.[41] Evidently, the interactive nature of the Japanese teaching style and the use of challenging problems are managed so well that the content is actually enhanced. We believe that a key reason for this high performance level is the efficient use of grappling and telling coupled with the benefits of disguised reinforcement exercises.

Additional analysis shows that 53 percent of the Japanese lessons used proof-based reasoning, whereas the comparable statistic for the U.S. lessons—which included both traditional and reform programs—stood at zero.[42] And in terms of the development of concepts and their depth and applicability as well as in terms of the coherence of the material, the quality assessments were much the same.[43] By all evidence, the use of proof-based reasoning as reported in Japan is not at all representative of the reform programs in the United States, and the use of such remarkably challenging problems seems beyond the scope of any U.S. program past or present.

When comparing U.S. reform practices and Japanese teaching methods, the Videotape Study offers somewhat guarded conclusions that are sometimes difficult to interpret. The report reads:

> Japanese teachers, in certain respects, come closer to implementing the spirit of current ideas advanced by U.S. reformers than do U.S. teachers. For example, Japanese lessons include high-level mathematics, a clear focus on thinking and problem solving, and an emphasis on students deriving alternative solution methods and explaining their thinking. In other respects, though, Japanese lessons do not follow such reform guidelines. They include more lecturing and demonstration than even the more traditional U.S. lessons [a practice frowned upon by reformers], and [contrary to specific recommendations made in the NCTM Professional Standards for Teaching Mathematics,[44]] we never observed calculators being used in a Japanese classroom.[45]

Subsequent elaboration on the similarities between U.S. reform and Japanese pedagogy recapitulates these ideas in the context of various reform goals, but again offers no statistical evidence to compare with the data accumulated from the analysis of Japanese teaching practices.[46] Consequently, it is difficult—absent additional context—to compare these reform notions in terms of mathematical coherence, depth, international grade level, or the preparation of students for more advanced studies and challenging problems. Not surprisingly, "the spirit of current reform ideas" seems difficult to measure. Similarly, the Japanese and U.S. reform pedagogies appear incomparable in their management of classroom time, their use of proof-based reasoning, their tradeoffs between student discovery and the use of grappling and telling, as well as their use of individual and small-group activities.

These distinctions notwithstanding, the notion that Japanese teaching might be comparable with U.S. reforms is given even greater emphasis in a major government report, which flatly declares: "Japanese teachers widely practice what the U.S. mathematics reform recommends, while U.S. teachers do so infrequently."[47]

This report on best teaching practices worldwide makes no mention of any differences between the U.S. reforms and Japanese teaching styles. Evidently, its perspective differs from that of its source of primary information, which is the more cautiously worded TIMSS Videotape Study.[48] Moreover, even the differences identified in the Videotape Study—which concern direct instruction, calculators, and teacher-managed demonstrations—are all matters of contention in the U.S. debate over classroom reform.

Lastly, it is significant (but seldom reported) that the Videotape Study makes a distinction between the idealized goals as prescribed in the NCTM Professional Standards for Teaching Mathematics and as embodied in actual classroom practices of some reform programs. In particular, the study discusses two reform-style lessons. One involves playing a game that happens to be devoid of mathematical content. The teacher claims this lesson is in accord with NCTM teaching standards. Stigler et al. disagree: "It is clear to us that the features this

teacher uses to define high-quality instruction can occur in the absence of deep mathematical engagement on the part of the students."[49] The other lesson was deemed to be compliant with the spirit of NCTM reforms. It began with the teacher whirling an airplane around on a string. The class then spent the period in groups exploring how to determine the speed of the plane and coming to realize that the key issues were the number of revolutions per second and the circumference of the plane's circular trajectory. The homework was a writing assignment: The students were asked to summarize their group's approach and to write about the role they played in the group's work. The study did not evaluate the content by grade level nor compare the lesson against the qualities that seem representative of Japanese teaching practices.

The Videotape Study reported that there was, apart from some minor differences, "little quantitative evidence that reform teachers in the United States differ much from those who claim not to be reformers. Most of the comparisons were not significant."[50] However, it is not evident how effective the study's comparison categories were at quantifying the key differences in various teaching practices.

Other Characterizations of Japanese Classroom Practices

Studies that use human interaction as a primary source of data must rely on large numbers of interpretations to transform raw, complex, occasionally ambiguous, and even seemingly inconsistent behavior into meaningful evidence. Given the complexity of the lessons, it is not surprising that different interpretations should arise. The Videotape Study—to its credit—documents an overview of these decision procedures, although their specific applications were far too numerous to publish in detail. Moreover, the study actually contains a wide diversity of observations, ideas, and conclusions, which sometimes get just occasional mention and are necessarily excluded from the Executive Summary. Understandably, this commentary—along with any supporting context—is also missing from the one-sentence to one-paragraph condensations in derivative policy papers.[51] Perhaps the seventh and eighth words in the opening line of the

study's Executive Summary explain this issue as succinctly as possible: "preliminary findings."[52] It is now appropriate to explore these larger-picture observations and to place them within the context of actual lessons.

The Videotape Study even offers a couple of sentences that support our own observations:

> [Japanese] students are given support and direction through the class discussion of the problem when it is posed (figure 50), through the summary explanations by the teacher (figure 47) after methods have been presented, through comments by the teacher that connect the current task with what students have studied in previous lessons or earlier in the same lesson (figure 80), and through the availability of a variety of mathematical materials and tools (figure 53).[53]

Unfortunately, these insights are located far from the referenced figures and the explanations that accompany them. The words are effectively lost among the suggestions to the contrary that dominate the report. It is also fair to suggest that the wording and context are too vague to offer any inkling of how powerful the "support and direction through class discussion" really was, and likewise, the value of the connections to previous lessons is left unexplored. This discussion does not even reveal if these connections were made before students were assigned to work on the challenge problems or after. For these questions, the video excerpts provide resounding answers: The students received masterful instruction.

The Videotape Study's Math Content Group analyzed thirty classroom lesson tables that were selected to be representative of the curriculum. Their assessments, as sampled in the study, agree with our overall observations, apart from the use of hints, which were mostly omitted from the lesson tables. Unfortunately, the analyses are highly stylized with abstract representations for use in statistical processing and were, presumably, not intended to be a reference for the actual teaching.[54]

Another sentence in the study begins with the potentially enlightening observation that:

> The teacher takes an active role in posing problems and helping students examine the advantages of different solution methods, *[however, rather than elaborating on how this takes place, the sentence changes direction with the words]* but the students are expected to struggle with the mathematical problems and invent their own methods.[55]

This interpretation of student work as inventive discovery appears throughout the TIMSS Videotape Study. In its analysis of the excerpted Japanese geometry lesson, the study categorizes the teacher's review of the basic solution method (shown in Figure 5.1) as "Applying Concepts in New Situation,"[56] but inexplicably switches tracks to count the student applications as invented student-generated alternative solution methods. Another such instance reads, "Students will struggle because they have not already acquired a procedure to solve the problem."[57] Similarly, the study never explains how teachers participate in the problem solving by teaching the use of methods and by supplying hints. Its only discussion about hinting is to acknowledge the offer of previously prepared hint cards.[58] And by the time the Glenn Commission finished its brief encapsulation of student progress, even the struggle had disappeared along with proper mention of extensive teacher-based assistance.

Searching for Answers

Let there be no doubt: The fact that we found no evidence of widespread inventiveness or student discovery should not be interpreted as a condemnation of exploration by students. Rather, it suggests a need for balance based on a realistic recognition of what can and cannot be done in classrooms.

Creativity and independent mathematical thought should be fostered, and alternative solution methods should be encouraged and studied. Students need to know that problems can be solved in different ways. They should learn to step back from a problem and think about plausible solution methods. And they need experience in selecting the best strategies for plans of first attack.[59] Similarly, students should learn firsthand how problems are

adapted to fit the method and how methods can accommodate new problems.

The Japanese lessons illustrate master instruction designed to foster this higher-level reasoning. When combined with modeling, these activities comprise the essence of problem solving.

However, despite the wealth of hints, the careful reviews of the necessary material, and the presumptive benefits accumulated from years of exposure to these teaching practices, the students discovered no new principles, theorems, or solution methods. And despite extensive assistance, many students did not conquer the first challenge problem of the day. These are sobering facts, and their implications for mathematics education should not be overlooked.

Just imagine: If the application of principles already learned and just reviewed is so difficult, consider how hard it must be to devise new principles. Ask mathematicians what they can do with three minutes of original thought. Chances are your answer will be no more than a quizzical look. New principles do not come cheap; research mathematics—even when there is strong evidence to suggest what might be true—requires enormous amounts of time. And eighth-graders will find the concepts and principles underlying eighth-grade-level math just about as difficult to develop. In short, there is a fundamental difference between problem solving and developing new principles. There are world-class mathematicians who are mediocre problem solvers and vice versa.[60] Few mathematical researchers would ever confuse the art of problem solving with the development of new mathematics. The implications for K–12 education and mathematics pedagogy are clear. Before we can understand what teachers and students should be doing in daily lessons, we must have a deep understanding of what they are doing as well as of what they can and cannot do. These distinctions—profound but sometimes subtle—lie at the heart of why modern mathematics developed over a period of two centuries or so and why arithmetic and elementary mathematics took even longer.

Conclusions

Large-scale video studies must rely on data coding and all kinds of preliminary judgments and filterings to encapsulate raw data. To cut through these sources of potential information loss and possible confusion, this study did something that the others did not. We supported our observations with a combination of the actual video images, a meticulous analysis of the mathematics lessons, and detailed citations together with a careful presentation of the context for each reference. Similarly, we sought to include relevant information regardless of whether or not it supported our conclusions. And whenever inconsistencies surfaced, we endeavored to reconcile the differences.

Of course, we must avoid extrapolating from a few "representative" tapings to draw conclusions about a much larger set of lessons, much less about the national characteristics of classroom teaching in the United States, Germany, and Japan. But with 229 lessons unavailable, and just six representative classes in view, there is little choice but to analyze the evidence that is in the public domain. Accordingly, this study should be viewed as a cautionary warning about widely cited opinions that might in fact be erroneous.

In summary:

- The videotapes of Japanese lessons document the teaching of mathematical content in a style that is deep and rich.
- The excerpts do not support the suggestion that in Japan, "[the] problem . . . comes first [and] . . . the student has . . . to invent his or her own solutions."[61]
- The evidence does suggest that in Japan, "students rarely work in small groups to solve problems until they have worked first by themselves."[62]
- Similarly, the evidence gives little weight to the notion that "Japanese teachers, in certain respects, come closer to implementing the spirit of current ideas advanced by U.S. reformers than do U.S. teachers."

- The evidence does confirm that "in other respects, Japanese lessons do not follow such reform guidelines. They include more lecturing and demonstration than even the more traditional U.S. lessons . . . ."[63]
- The excerpts show Japanese classes featuring a finely timed series of minilessons that alternate between grappling-motivated instruction on how to apply solution methods and well-chosen challenge exercises designed to instill a deep understanding of the solution methods just reviewed. No other interpretation is possible.
- Some official U.S. government reports overemphasize unsubstantiated claims about Japanese pedagogy while omitting all mention of the remarkably high-quality *instruction* that is characteristic of Japanese lessons.
- Studies of problem solving in the classroom should include statistical analyses of as large a variety of practices and interactions as possible, including the use of grappling and telling, in-progress hints and mentoring, and preparatory discussion with hints and applicable content. Similarly, the roles of teacher assistance in presentations of all kinds ought to be better understood.
- Research projects in mathematics education should strive to maintain open data to support independent analyses. In addition, great care should be exercised to ensure that the encodings and analyses incur no loss of mathematical content or pedagogy.

It is perhaps fitting to close with a few words that strip away the citations, figures, tables, and video images that characterize the preceding analysis and to express some observations in more human terms.

Everyone understands that students must learn how to reason mathematically. The heart of the matter, therefore, is how—not whether—to teach problem solving and mathematical investigation. We must not be so desperate for the teaching of problem

solving that we acclaim all such efforts to be one and the same and, therefore, equally promising. The video excerpts document exemplary instances of master teachers instructing students in the art of adapting fundamental principles to solve problems. In each sample excerpt, the class had already learned the basic method necessary to solve the challenge problems of the day. However, students had to possess a solid understanding of the method before it could be applied successfully.

This form of teaching requires a deep understanding of the underlying mathematics and its difficulty. Students must be properly prepared so that they can master the content at an adequate pace. Whenever hints are necessary, the teacher must be sensitive to these needs and stand ready to offer whatever assistance is appropriate to open the eyes of each individual learner. More often than not, most students will be unable to apply fundamental principles in new settings until they see step-by-step examples completed by the teacher. In these cases, the students should then get the opportunity to walk in the teacher's footsteps by applying the approach to a new problem that is designed to have the same challenges in a slightly different context.

These are the lessons that must be learned from the videotape of Japanese teaching. As the excerpts demonstrate, a master teacher can present every step of a solution without divulging the answer and can, by so doing, help students learn to think deeply. In such circumstances, the notion that students might have discovered the ideas on their own becomes an enticing mix of illusion intertwined with threads of truth. Unfortunately, such misunderstanding risks serious consequences if it escalates to a level that influences classroom practice and education policy. In retrospect, it seems appropriate to offer one last cautionary recommendation: Unless lesson studies include a comprehensive analysis of the mathematics content and the full range of teaching techniques, their conclusions will perforce be incomplete and, as a consequence, vulnerable to misconceptions about the very practices that best enhance student learning.

Notes

1. Cf. J. W. Stigler et al., *The TIMSS Videotape Classroom Study: Methods and Findings from an Exploratory Research Project on Eighth-Grade Mathematics Instruction in Germany, Japan, and the United States* (National Center for Education Statistics, 1999), 134.
2. J. Glenn et al., *Before It's Too Late: A Report to the Nation from the National Commission on Mathematics and Science Teaching for the 21st Century,* Report EE0449P (Education Publications Center, U.S. Department of Education, September 27, 2000), 16.
3. M. O. Martin et al., *School Contexts for Learning and Instruction: IEA's Third International Mathematics and Science Study* (TIMSS International Study Center [ISC], 1999).
4. I. V. S. Mullis et al., *TIMSS 1999 International Mathematics Report, Findings from the IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade* (TIMSS ISC, Dec. 2000).
5. J. W. Stigler et al., *TIMSS Videotape Classroom Study.*
6. J. W. Stigler and J. Hiebert, *The Teaching Gap: Best Ideas from the World's Teachers for Improving Education in the Classroom* (Free Press, 1999).
7. *Eighth-Grade Mathematics Lessons: United States, Japan, and Germany* (Videotape, NCES, 1997).
8. *Moderator's Guide to Eighth-Grade Mathematics Lessons: United States, Japan, and Germany* (NCES, 1997).
9. Video Examples from the *TIMSS Videotape Classroom Study: Eighth-Grade Mathematics in Germany, Japan, and the United States* (CD ROM, NCES, 1998).
10. In Figure 5.1, the translation shows an "or" instead of an "and." This mathematical error is due to a mistranslation of the spoken Japanese.
11. *Moderator's Guide,* 136.
12. Ibid., 140.
13. Ibid.
14. Ibid., 139–41.
15. Ibid., 141.
16. The assignment probably should be restricted to convex figures; otherwise it includes irregular cases that are difficult to formalize. On the other hand, this concern is just a minor technicality that has no effect on the pedagogical value of the problem.
17. In fact, the technique is central to Euclid's development of area in general, which is based on transforming any polygon into a square with the same area. And the nat-

ural extension of this problem became a question for the ages: how to square the circle.

18. *Moderator's Guide,* 114.
19. Ibid.
20. Ibid., 159.
21. Ibid., 164.
22. J. Glenn et al., *Before It's Too Late,* 4.
23. J. W. Stigler et al., *TIMSS Videotape Classroom Study,* vi.
24. Ibid., 55.
25. Ibid., 97.
26. Ibid., 26–27.
27. In particular, the *Moderator's Guide* (pages 161–63) discusses this one unsuccessful approach as the entirety of the section titled "Students Presenting Solution Methods."
28. J. W. Stigler et al., *TIMSS Videotape Classroom Study,* 24.
29. Ibid., 31.
30. Ibid., 100.
31. Cf. J. W. Stigler et al., *TIMSS Videotape Classroom Study,* vi; L. Peak et al., *Pursuing Excellence: A Study of U.S. Eighth-Grade Mathematics and Science Teaching, Learning, Curriculum, and Achievement in International Context* (NCES, 1996), 9; and J. Glenn et al., *Before It's Too Late,* 16.
32. J. W. Stigler and J. Hiebert, *The Teaching Gap,* 79.
33. J. W. Stigler et al., *TIMSS Videotape Classroom Study,* 134.
34. Ibid., 78.
35. A. E. Beaton et al., *Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study* (TIMSS ISC, 1996), 154–55.
36. P. Cobb, E. Yackel, and T. Wood, "A Constructivist Alternative to the Representational View of Mind in Mathematics Education," *Journal for Research in Mathematics Education* 23 (1992): 28.
37. D. L. Schwartz and J. D. Bransford, "A Time for Telling," *Cognition and Instruction* 16 no. 4, (1998): 475–522.
38. Cf. D. L. Schwartz and J. D. Bransford, "A Time for Telling," and J. D. Bransford et al., *How People Learn: Brain, Mind, Experience, and School* (National Research Council, National Academy Press, 2000), 11.
39. J. W. Stigler and J. Hiebert, *The Teaching Gap,* 48–51.
40. Ibid., 49.
41. J. W. Stigler et al., *TIMSS Videotape Classroom Study,* 44.

42. Ibid., vii.
43. J. W. Stigler and J. Hiebert, *The Teaching Gap,* 59.
44. The bracketed additions are elaborations from page 123 of the *Videotape Study,* where the discussion of calculator usage is reworded and thereby avoids the grammatical misconstruction we have caused with the unedited in-place insertion.
45. J. W. Stigler et al., *TIMSS Videotape Classroom Study,* vii.
46. Ibid., 122–24.
47. L. Peak et al., *Pursuing Excellence,* 9. See also pages 41 and 43.
48. J. W. Stigler et al., *TIMSS Videotape Classroom Study.*
49. Ibid., 129.
50. Ibid., 125.
51. Cf. L. Peak et al., *Pursuing Excellence,* and J. Glenn et al., *Before It's Too Late.*
52. J. W. Stigler et al., *TIMSS Videotape Classroom Study,* v.
53. Ibid., 134.
54. Ibid., 58–69. For example, the analysis of the excerpted geometry lesson consists of a directed graph with three nodes, two links, and nine attributes. The first node represents the basic principle (attribute PPD) for the presentation illustrated in Figure 5.1. The node's link has the attributes NR (Necessary Result) and C+ (Increased Complexity). It points to a node representing the Eda-Azusa challenge exercise. The representations were used to get a statistical sense of various broad-brush characteristics of the lessons.
55. Ibid., 136.
56. Ibid., Figure 63, 101.
57. Ibid., 35.
58. Ibid., 26–30.
59. It is worth noting that the German algebra lesson (unlike either of the U.S. lessons) also covered strategy. The excerpted lesson on two equations in two unknowns has a review of the three solution methods that had been already taught. Then a more difficult problem that has two additional features is introduced. First, it requires the collection of like terms. Second, the coefficients permit the solution methods to be applied to one of the variables more easily than the other. This second issue seems to have been missed by the entire class and is revealed by the teacher only after the class has worked (too hard) to solve the problem. There is also some discussion about the advantages and disadvantages of each solution method.
60. Of course, problem solving is one component of research mathematics, but it can have a remarkably minor role in the very complex art of formalizing and establishing mathematical frameworks and fundamental principles.

61. Cf. J. W. Stigler et al., *TIMSS Videotape Classroom Study,* vi.
62. J. W. Stigler and J. Hiebert, *The Teaching Gap,* 79.
63. J. W. Stigler et al., *TIMSS Videotape Classroom Study,* vii.