CHAPTER ONE

Assessments in an RTI System

Fulfilling the promise of response to intervention (RTI) demands balance. Our idealistic notion of proactive, adaptive instruction must be weighed against the realities of schools and classrooms. In considering the detailed information offered by present-day assessments, we must weigh the time they require against the benefits we can expect from having given them. Although we adhere to the idea that assessments can guide our efforts to plan targeted instruction, we do not believe that more is always better. We have discovered that the benefits of RTI can be realized with no more than a simple set of informal assessments applied strategically.

That simple set of assessments must be chosen to help reach the following goals:

1. Quickly screen all students to determine areas of difficulty.
2. Follow up with diagnostic measures to help plan targeted instruction.
3. Periodically monitor progress to determine the near-term impact of that instruction.
4. Collectively aid in determining next steps.

Using assessments to achieve these goals requires a solid understanding of how they work and what they can tell us. It also requires a system that ensures their sparing and deliberate use, governed by a decision-making strategy that is clear to all.

THE ASSESSMENT DILEMMA: FINDING THE "JUST RIGHT" AMOUNT

The poet Robert Browning once offered this advice to painters: "Less is more." By that he meant that excessive detail often serves little purpose. We believe that, up

to a point, this idea can be applied to reading assessment. Imagine a school where every child was given a full clinical battery of tests. Information would abound, and in a very few cases nearly all of it would be useful in planning interventions. However, for the vast majority of students, most of the information would reveal few unique insights essential for guiding effective classroom instruction. In our two university clinics, we occasionally encounter such students. They are referred by frustrated teachers or anxious parents, but they actually present simple profiles that extensive diagnostic assessments only confirm. In short, they are overtested.

An efficient RTI assessment system is lean and mean. It embraces a "Goldilocks" approach in which just the right amount of assessment is conducted to maximize student growth. The extremes are avoided. Too little assessment can result in vague guidance for teachers and instruction that is not sufficiently targeted. Too much assessment rarely results in those "aha moments" that provide the key to student success. Overassessing also requires time that might have been devoted to instruction. After all, during an assessment, the student is not learning and the teacher is not teaching. Unless the results offer practical insights into the kind of instruction that will serve the student best, this is a lose–lose situation.

Figure 1.1 shows the relationship between the amount of assessment we conduct and the amount of student growth we can expect on the basis of what we learn. This is not an empirical graph, but it represents, in a general way, our combined experience in classroom and clinic. In the early stages of assessment, the information we obtain about a child can be used to plan instruction that is likely to be far more effective than what we might have provided before the assessment. For example, determining an appropriate level of text would help immensely in placing the child in materials that will allow optimal progress. However, in giving additional assessments, we soon reach a point at which less and less useful information

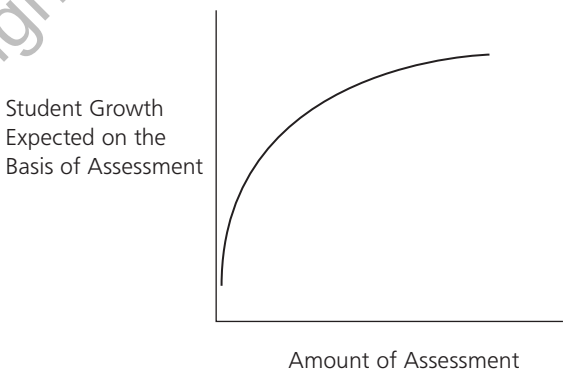

FIGURE 1.1. The relationship between the amount of assessment and student growth.

is obtained. The key is to find the point where we know enough to plan effective instruction.

THE ROLES OF ASSESSMENT IN RTI

If we are to conduct the "just right" amount of assessment—no less, no more—we need a system to direct our efforts. Of central importance in developing such a system is an understanding that assessments are of different kinds and serve a variety of purposes. Like tools, assessments are designed for specific functions. Using them for other purposes can be misleading and counterproductive. Unless you are like McKenna, for example, you would never use a wrench to do the job of a hammer.

Types of Assessments

To ensure a good understanding, let's begin with an overview of the principal kinds of assessments and how they might be used in an RTI system. From there, we'll review the uses of these assessments.

Norm-Referenced versus Criterion-Referenced Assessments

Test scores must be interpreted in order to make them useful. A score alone, without a frame of reference for interpreting it, has no meaning. As an example, consider an informal test designed to determine whether a child can apply the rule of silent *e* in decoding one-syllable words. Such a measure is one subtest of the Informal Phonics Inventory (Form 5.3 of *Assessment for Reading Instruction*, Second Edition [McKenna & Stahl, 2009]). The child views the following pairs of words:

cap	tot	cub	kit
cape	tote	cube	kite

For each pair, the teacher points to the upper word and asks the child to pronounce the lower one. ("If this is *cap*, what is this?") Scores on this subtest can obviously range from 0 to 4. But what does the score for a given child tell us? One way to give it meaning is to use a cutoff score, or criterion, to help us judge whether the skill assessed has been mastered. If we are satisfied with the criterion, the score can help us determine whether instruction in the rule of silent *e* is desirable. Another way to give meaning to the score would be to compare it to the scores of other children. Because this skill is typically taught in first grade, we could judge the score in terms of what is normal at a particular point in time. These two basic approaches

to interpretation lead to different conclusions about test performance and they are intended to answer different questions.

CRITERION-REFERENCED TESTS

When the goal of assessment is to determine whether a skill has been mastered, a criterion-referenced assessment can be useful. For example, interpreting that portion of the informal phonics inventory that assesses the rule of silent *e* has an 80% criterion associated with it. If a student scores at or above this level, a teacher is justified in concluding that additional, targeted instruction in this skill is not required. Note that because only four items are administered, a perfect score is needed to denote mastery.

Criterion scores are useful in two cases. When assessing skills that are constrained (i.e., skills for which total mastery is possible), criterion scores can help determine whether or not mastery has been attained. The silent *e* subtest is a good illustration. When a child meets the criterion, the issue is settled. The second case involves a skill that is never totally mastered but for which we can establish criteria for specific points in time. We call such criteria benchmarks because they are determined through longitudinal studies designed to predict future performance. Tests of oral reading fluency, such as those included in Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Next and AIMSweb, are examples of assessments for which shifting benchmarks provide useful gauges of a student's needs. The DIBELS Next benchmark for the Oral Reading Fluency (ORF) subtest in the fall of grade 2 is 53 words correct per minute (WCPM) for passages written at a high second-grade level. But in the fall of grade 5, the benchmark is 111 WCPM for passages written at a high fifth-grade level. Developmental shifts and text demands influence the formation of benchmarks for commercially produced tests.

To sum up, criterion scores can serve either as indicators of a child's mastery of constrained skills or as benchmarks that rise with time and task. In both cases, the measurement issue is between the child and the skill. The performance of other children on the assessment is not directly involved in interpreting the score.

NORM-REFERENCED TESTS

When the goal of assessment is to compare a child with the overall population of children, a norm-referenced test is appropriate. Here, a child's raw score is converted into one or more norms, which are converted scores used to make comparisons possible. Many norms are possible, but only a few are typically used in RTI assessments. Three of the most common are defined in Table 1.1, along with their strengths and weaknesses.

Now let's consider the example of oral reading fluency from a normative standpoint. According to the norms developed by Hasbrouck and Tindal (2006), a

TABLE 1.1. Characteristics of Common Norms

Norm	Definition	Advantages	Drawbacks
Percentile rank	Percentage of age peers that a child's score equals or exceeds	• Relatively fine grained	• Cannot be averaged • Are not linear, making gains hard to interpret
Stanine	One of nine statistically equivalent categories, with one lowest, five average, nine highest	• Ease of comparison, using two-stanine rule to judge significant differences	• This gross, at-a-glance measure may hide small differences and gains
Grade equivalent	Estimated grade and month associated with a test score	• Appropriate for some adaptive tests	• Easily misinterpreted • Usually computed by extrapolation rather than by assessing children at various grades • Discouraged by the International Reading Association

beginning second grader who reads 53 WCPM would score at the 50th percentile rank. This means that the child is exactly in the middle of the pack, dead average. This information provides a second frame of reference by which to judge performance based on that of other children; when available it can be considered in tandem with a benchmark (a cutoff score predictive of future success).

Curriculum-Based Measures

A curriculum-based measure (CBM) is a type of standardized test that is aligned with grade-level curriculum. The original CBM tasks were actually constructed using samples of a school's curriculum materials. However, today's CBMs are produced commercially to reflect different components of a grade-level curriculum area. The commercially produced CBMs are most popular because the resources invested in mass production increase the ability to provide tests that are *technically adequate*, meaning that they are valid and reliable. The tests are usually timed, enabling them to be sensitive to small margins of growth. Data from CBMs are easily summarized on charts and within web-based data management systems.

GENERAL OUTCOME MEASURES

General outcome measures (GOMs) are a type of CBM that assess the general outcome on a complex task that is not divided into subskills. For example, oral reading fluency and maze are designed to be GOMs of fluency and comprehension, respectively. Children are given reading passages that typically developing children would be expected to be able to read at the end of a particular grade level. GOMs are a

means of looking at students performing a complex task when looking at individual subskills does not really portray the desired overall instructional goal. In order to read fluently, a number of foundational subskills must be operating together. GOMs assess this overall operation. However, if children are having performance difficulties, the GOM does not provide specific diagnostic information because it is designed to look at general overall performance on a capstone task.

SKILL-BASED MEASURES

Skill-based measures (SBMs) are similar to GOMs in that they measure sets of individual skills that are likely to be accomplished by the end of a school year. However, rather than the assessment requiring a single process that requires the interaction of multiple skills, each SBM is composed of mixed items from a set of goals. This type of assessment is commonly used in math to measure growth on individual computational skills across a school year. Each SBM might consist of a random collection of each type of computation that children are expected to master by the end of the year. As children improve over time, the total scores go up. However, items are keyed to particular skill areas and can be used to inform instruction.

MASTERY MEASURES

Mastery measures (MMs) are CBMs that measure discrete skills. For example, those CBMs that measure letter identification fluency or letter–sound fluency are considered MMs. They are useful when it is important to monitor a skill that is taught in isolation or while troubleshooting a particular area that is giving a student difficulty.

Functions of Assessment

In order for RTI to work, the assessments must serve a variety of purposes. They must quickly identify the existence of problems. They must help to identify specific deficits to be targeted through instruction. They must reveal whether students are responding to targeted instruction. Finally, they must produce long-term results useful in evaluation at the level of the classroom, the grade level, the school, and the district. These four functions are usually labeled as screening, diagnostic, progress monitoring, and evaluation.

Screening

Screening assessments are universally administered in an RTI system. They are either quick to administer (such as an AIMSweb or DIBELS fluency passage) or they already exist (such as an end-of-year achievement test from the previous year).

Because all students must be assessed, screening tests reveal only broad portraits of individual students. They lack sufficient detail to plan instruction. Their chief advantage lies in identifying students who are experiencing problems in a particular area. For those identified, a more fine-grained assessment is required, one designed to provide diagnostic information. Essentially, a trade-off is involved. What we gain in speed and efficiency, we lose in the specific information needed to plan instruction. Consider the DIBELS Next subtest Nonsense Word Fluency (NWF). In this test, a child pronounces rows of one-syllable pseudowords and is halted after a minute. Comparing the raw score (the sum of letter-sounds pronounced) against a benchmark can help determine whether basic decoding is a problem area. What it cannot do is identify specific skill deficits a teacher should address through instruction. We disagree with those who advocate inspecting individual nonsense words in an effort to determine these deficits (e.g., Hall, 2006). This is because the letter-sounds that make up the words are not presented systematically and also because the child's success or failure will depend in part on adjacent letter-sounds. For example, if a child pronounces *niz* by saying only the initial consonant, we cannot be certain that he or she has no knowledge of the sound /z/. This means that a follow-up assessment is needed, one that is designed to serve a diagnostic function.

Diagnostic

Diagnostic assessments provide information about a problem area in sufficient detail that targeted lessons can be planned. Because of the time they require to administer and score, they are not administered universally but only when screening has indicated a problem. Diagnostic testing is associated with a number of misconceptions. Teachers often believe that they are commercial tests that "come in a box," and are so involved that only specialists can administer and interpret them. Although a few diagnostic assessments are like that, the type most useful in RTI is informal and easy to give. Such a test yields information that is immediately useful in planning instruction. As one example, consider the Informal Phonics Inventory (McKenna & Stahl, 2009). After DIBELS NWF has indicated that a child is below the benchmark, the inventory can identify specific deficits. Because of the time required to give the inventory and because of the fact that many students do not exhibit problems with decoding, it is neither practical nor necessary to administer it to all students.

Assessments designed to diagnose rarely make good screeners. However, they are useful for targeting instruction. Without diagnostics, it would be difficult for teachers to plan deliberate interventions that can meet students' needs in the most time-efficient manner. Without diagnostics, instruction may address the general need without providing the specificity that facilitates accelerated growth. Skilled interpretation of the diagnostics allows teachers to teach only the content that is needed and to skip the instructional content that is already mastered or content that

is beyond the students' zone of proximal development (ZPD)—that which they can do with assistance.

Progress Monitoring

Progress monitoring assessments are the mainstay of RTI. They are given periodically to determine whether a child is responding to the intervention provided. They are often alternate forms of the same tasks used as screening tests. Wixson and Valencia (2011) identify two types of progress monitoring tests: (1) formative tests, used to gather information while instruction is under way, and (2) summative tests, which are given for benchmarking purposes, typically in fall, at midyear, and in spring. Monitoring progress allows teachers to know when their instruction is working and when a course change is required. Progress monitoring scores can be recorded as a record of an individual's growth over time, and this record can be instrumental in deciding whether more intensive instruction is needed. These scores can also be averaged at the classroom, grade, and school levels to track success over time and from grade to grade. Assessments useful in progress monitoring include alternate forms of standardized instruments, such as those available from DIBELS and AIMSweb, but they can also include running records and teacher-constructed measures geared precisely to the content taught.

Evaluation

Evaluation assessments are aimed at determining whether teachers and schools are meeting the collective needs of students. The most prominent examples are undoubtedly the high-stakes assessments required by No Child Left Behind and other outcome tests. These tests are the "bottom line" of RTI and they serve the interests of some stakeholders. However, they tell us relatively little about individual students beyond a tentative screening function. More sensitive indicators of the health of an RTI program lie in the screening and progress monitoring measures. When scores are examined collectively, the "state of the school" can be described in considerable detail (Walpole & McKenna, 2012). For example, the percentage of children at risk should fall throughout the year if they are responding to the intervention they receive. Scores on screening and progress monitoring measures can be examined by teacher and grade level to help coaches identify priorities. We return to this idea in Chapter 8.

Function versus Type of Assessment

A frequent confusion about the basic functions of assessments is the belief that a particular test can have only one function. The problem with this belief is that the same test can often serve several functions. Table 1.2 lists examples of familiar tests

TABLE 1.2. Examples of Familiar Tests and the Functions They Can Serve

Test	Screening	Diagnostic	Progress monitoring	Outcome
DIBELS Next	✓		✓	✓
AIMSweb	✓		✓	✓
Informal Phonics Inventory		✓	✓	✓
Inventory of High-Frequency Words		✓	✓	✓
Running Record of Text Reading (K–2)	✓	✓	✓	
Group Achievement Test	✓			✓

and the functions they can serve. Although we might argue about whether a certain test can serve a certain function, there can be no disputing that many (perhaps all) tests can serve more than one.

Using Tests in Tandem

Using screeners to identify problem areas and diagnostics to narrow those areas to practical instructional targets is the great one–two punch of assessment-driven instruction (McKenna & Walpole, 2005). Once targeted instruction begins, progress monitoring tests come into play, helping us gauge the extent to which the instruction is having the desired effect. Figure 1.2 illustrates the decision-making process that is guided by the results of all three types of assessments. Note the return loop back to targeted instruction when progress monitoring indicates that the problem persists. This pathway is central to RTI, for it reflects an awareness that a child is not responding and requires that targeted instruction be reconsidered.

Summative and Formative Uses of Tests

The terms *summative* and *formative* are an occasional source of confusion. The difference lies in how the results are used. Formative assessments yield results that are used to modify instruction. Progress monitoring is a type of formative assessment because the results may cause a teacher to alter an approach or to change course entirely. Summative assessments, on the other hand, are used to ground judgments after the fact. They may result in big-picture changes, such as whether a particular intervention program is effective or whether special education staffing is called for, but they are not used to make day-to-day judgments about what kind of instruction to provide. A comparison with cooking is sometimes used to explain the difference:

> *When the cook tastes the soup, that's formative.*
> *When the guest tastes the soup, that's summative.*

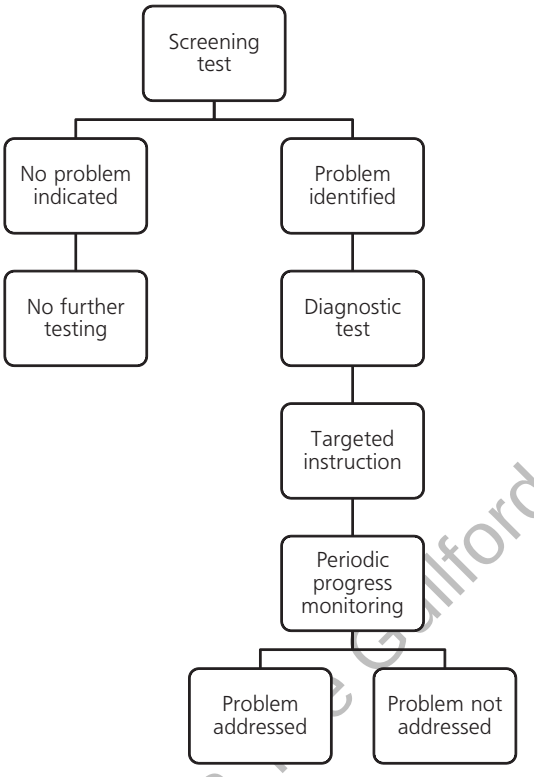

FIGURE 1.2. How different assessments are used to target instruction.

ENSURING FIDELITY

The elegance of an RTI assessment system lies in using the least amount of testing to make prudent decisions for every child. Making sound decisions based on a small number of tests requires that they be administered, scored, and interpreted properly. Every test involves a small amount of measurement error, and our goal must be to keep that error as small as possible. When guidelines are not followed, we increase the magnitude of the error and bad decisions can result. To illustrate how important the three dimensions of fidelity can be, let's consider the example of the DIBELS ORF subtest.

Fidelity to Administering an Assessment

Fidelity in administering the ORF requires that three passages be administered, that the child be stopped 1 minute into each passage, and that only the middle score

be recorded. We have known teachers to take the tempting shortcut of giving only one passage. In doing so, they are gambling that the passage they choose is representative of the child's actual fluency level. On the ORF the one passage selected may be higher or lower than the child's true score. Although it is true that none of the three scores is likely to be a perfect reflection of the child's proficiency, we can have greater confidence in it because it fell between two other scores. Consider a beginning fourth grader who takes one DIBELS passage and scores 78. The teacher correctly judges the child to be performing below the benchmark (90). But what if the teacher were to administer two more passages, as the instructions indicate, and the child scores 90 and 95? In that case, the middle score would have been at benchmark and the teacher would have reached the wrong conclusion.

Another example involves gaming the test to improve the apparent level of performance. We have known a few teachers, under pressure to produce results, who have encouraged children to skip unfamiliar words during ORF testing. Because only the number of words correctly pronounced are counted—and not the errors—directing students to skip words rather than "lose time" attempting to decode them, leads to inflated scores. Such scores can hardly be the basis for sound decisions regarding the kind of instruction children should receive.

Fidelity to Scoring an Assessment

Fidelity to scoring is obviously important as well, but there are many ways to go wrong. On the ORF, the teacher might miscount the number of words attempted, undoubtedly the most common mistake. It is also possible that synonyms and other semantically acceptable substitutions might be counted correct (saying *dog* for *pup*). This practice seems reasonable and was once supported by some theorists, but it has now been shown to be misleading (McKenna & Picard, 2006/2007). More important, it was not used to determine the DIBELS benchmarks and can only inflate a child's score, leading to overestimates of proficiency. The teacher might also count all of the words attempted up to the 1-minute mark, including errors. Doing so would result in a measure of rate (WPM) rather than in the combined measure of rate and accuracy (WCPM). Rate alone is a limited and outdated measure of fluency. And once again, the DIBELS benchmarks are based on WCPM, which requires scoring on this basis alone. Yet another threat to fidelity is the temptation to give the benefit of the doubt. We all want our students to do well, but when we catch ourselves saying, "He's just having a bad day—I know he knows that word and I won't count it wrong," we are jeopardizing the results of the assessment.

In short, scoring fidelity can be compromised by a number of factors, some accidental, some deliberate. Everyone makes mistakes, to be sure, but teachers can minimize them by carefully reviewing scoring procedures and adhering to them. There are good reasons for doing so.

Fidelity to Interpreting an Assessment

Fidelity to interpretation depends on the nature of the test. For criterion-referenced tests, strict application of the mastery criteria or benchmarks is important. We don't mean to suggest that there is no room for professional judgment, and we have already acknowledged the part played by measurement error. It is prudent to think of a "gray area" just below the criterion and to use other information about a child's performance to interpret scores falling there.

In the case of a mastery test, such as those that make up the Informal Phonics Inventory, we have set the criterion at 80% but have created a zone just below it to indicate partial mastery. A child's performance can be categorized as follows:

Mastery	80–100%
Review	60–79%
Systematic Instruction	Below 60%

The "review" category creates a gray zone, created to avoid the all-or-nothing judgment that a mastery criterion implies.

In the case of a benchmark test, like ORF, there is also a gray area. Imagine that our beginning fourth grader had a middle score of 88. The benchmark for the beginning of fourth grade is 90, and technically the child is below benchmark. Is further consideration warranted? This is a judgment call for teachers seeking to form flexible groups for targeted instruction. DIBELS approaches the issue by establishing three levels of risk, often color coded as green, yellow, and red. This child would be placed in the yellow zone on computer-generated reports, but it is still left to the teacher to decide on an instructional course of action.

For norm-referenced tests, careful consideration of what the norms actually mean is required. Unlike a criterion-referenced test, which comes with well-defined scores for categorizing a child's performance, a norm-referenced test is more complicated. As we have said, there are several norms from which to choose, and each comes with unique methods of determining the gray area. Fortunately, however, such tests are not a major component of RTI. Their chief utility lies in end-of-year outcome assessments, such as nationally normed group achievement measures, and in the assessments given by special educators to determine the appropriateness of a particular category (e.g., learning disabled).

Guilford Publications
72 Spring Street
New York, NY 10012
212-431-9800
800-365-7006
www.guilford.com