CHAPTER 2

School-Based Assessment

There are several key elements to RTI, but effective assessment is at the very core of any successful implementation model. In fact, we endorse a previous definition of RTI (Burns & VanDerHeyden, 2006) that conceptualized RTI as the systematic use of assessment data to most efficiently allocate resources in order to enhance learning for all students. Thus the primary purpose of assessment within an RTI model is to facilitate instructionally relevant data-based decision making. Data within an RTI model are used to identify the need for an intervention, identify which intervention is mostly likely to lead to success, determine whether an intervention resulted in adequate response, and in some cases, decide whether special education services are warranted (Ysseldyke, Burns, Scholin, & Parker, 2010). In this chapter we discuss assessment in general and the types of assessments that can most adequately inform an RTI model.

The primary purpose of assessment within an RTI model is to facilitate instructionally relevant data-based decision making.

ASSESSMENT

The term *assessment* is used every day by practitioners from various fields and with varying connotations. We endorse the definition by the American Educational Research Association, American Psychological Association, and National Council for Measurement in Education (1999) that describes assessment as a decision-making process. Many terms that are used synonymously with assessment (e.g., testing) are actually potential components of an assessment process and do not have the same meaning. Specifically, assessment is an ongoing process of gathering information about student progress. Tests are one method to gather relevant data. Assessment processes lead to decisions that may or may not be valid, and the validity of the decisions must be evaluated within the purpose for which the data were gathered. There are no bad tests, just inappropriate uses of the data. Validity refers to the

degree to which an assessment measures what it claims to measure, and no assessment tool or procedure is valid "for all purposes or in the abstract" (Sattler, 2001, p. 115). We do not discuss basic psychometric issues of reliability and validity, but instead apply the concepts to an RTI model.

Burns, Jacob, and Wagner (2008) reviewed research and assessment standards to conclude that an RTI process can lead to decisions that are fair, valid, comprehensive, multifaceted, and useful if the protocols within the model are carefully crafted, and interventions are based on a scientific problem-solving process that involves identifying and clarifying the problem, generating solutions, and measuring outcomes. Moreover, RTI models should rely on assessments for which research has consistently demonstrated instructional utility. Most assessment tools used in schools today are inconsistent with the assessment purposes within an RTI framework. For example, end-of-the-year tests are typically not useful for making intervention decisions. Moreover, problem analysis and school-based assessments have focused primarily on collecting historical data to identify internal and unalterable student characteristics. If the goal is to change student behavior or trajectory of learning, then estimates of previous behavior serve little purpose except to establish a baseline. However, when previous measures of student behavior are used to identify the environmental conditions that created or maintained the problem behavior in order to generate testable hypotheses, and those hypotheses are tested with ongoing data collection, then positive outcomes are more likely to occur. In other words, assessments within an RTI framework fit within Stiggins's (2005) model of assessment *for* learning rather than assessment *of* learning.

FORMATIVE EVALUATION

There are several ways to categorize various uses and formats of assessment, but we focus on the formative–summative continuum because this is the most central issue to an RTI framework. Bloom, Hastings, and Madaus (1971) delineated two purposes of assessment: formative evaluation: (1) "systematic evaluation in the process of curriculum construction, teaching, and learning for the purposes of improving any of these three processes," and (2) summative evaluation—the collection of data after instruction occurred to make judgments about the instruction such as "grading, certification, evaluation of progress, or research on effectiveness" (p. 117). The primary goal of summative evaluation is to determine how much has been learned or how much is being learned, but formative evaluation suggests specific objectives and items that need to be taught and how best to teach them (Stiggins, 2005).

Formative evaluation procedures are critical for improving student outcomes and are essential to effective RTI practice, and are probably best accomplished with samples of student behavior before, during, and after interventions occur. Monitoring student progress, usually with curriculum-based measurement (CBM), has become synonymous with formative evaluation (Deno, 2003; Silberglitt & Hintze, 2005), but those data are collected during or after interventions to determine their effectiveness, which is the very definition of summative evaluation (Bloom et al., 1971). Monitoring progress is an important aspect of an RTI framework, but if practitioners are interested in implementing a formative evaluation

framework, which is needed for successful RTI implementation, then monitoring student progress is only scratching the surface.

Meta-analytic research found that most data collected before interventions, including IQ tests and standardized measures of reading, correlate quite well with pre- and post-intervention reading scores, but have minimal correlation with actual growth during the intervention (Burns & Scholin, in press). In other words, most standardized measures of IQ and reading predict who is a good reader and who needs help, but they do not predict for whom an intervention will be successful. Moreover, several researchers have suggested that practitioners should use data obtained from measures of various cognitive processes in order to determine appropriate reading interventions (Feifer, 2008; Fiorello, Hale, & Snyder, 2006; Hale, Fiorello, Bertin, & Sherman, 2003; Hale, Fiorello, Kavanagh, Hoeppner, & Gaither, 2001), but meta-analytic research found very small effects ($d \approx 0.20$) for interventions derived from measures of auditory or visual association, reception, and sequential memory (Kavale & Forness, 1999), and a recent meta-analysis found negligible effects for interventions derived from cognitive processing data ($g = 0.09$ to 0.17; Burns, Kanive, & Degrande,, 2012). Thus practitioners who use measures of cognitive processing as part of the intervention process are likely engaging in an ineffective practice based more on a resilient belief system than on research.

Although measures of cognitive processing do not suggest effective interventions because they lead to small effects, formative evaluation led to an average effect size of 0.71 (Fuchs & Fuchs, 1986), which suggested an effective practice. Why is formative evaluation so effective? Formative evaluation is characterized by data collected before instruction occurs (Linn & Gronlund, 2000), that are used to identify student needs and to plan instruction to better meet those needs (William, 2006). Thus formative evaluation should identify what to teach and how to teach it, which is probably best accomplished by direct samples of student behavior. This process results in the 0.71 effect size, suggesting effective practice. Below we discuss the characteristics of data that can be used to inform intervention and how those formative data fit within the purposes of an RTI model.

ASSESSMENT DATA FOR RESPONSE-TO-INTERVENTION DECISIONS

Ysseldyke and colleagues (2010) suggested that in order for data to be considered within an intervention process, they should be evaluated for their precision, potential frequency of use, and sensitivity to change. However, these are not absolute terms because different levels of precisions, frequency, and sensitivity are needed depending on how the data are used. Remember, there is no such thing as a bad test or assessment—there are only misappropriate uses of the data. For example, if a measure was designed to screen students (i.e., identify students who need additional support), then the data would likely not be useful for designing interventions. Alternatively, some data are well designed to identify specific areas of student need, but do not offer reliable or global enough data to provide a screening

of overall skill. Thus assessment tools must be evaluated within the context of how they are used.

Chapter 3 talks more specifically about uses of data to analyze problems, but here we discuss data that are used to (1) identify the need for an intervention, (2) identify which intervention is mostly likely to lead to success, (3) determine whether an intervention resulted in adequate response, and (4) decide whether special education services are warranted.

Identify Need for Intervention

The first decision within an RTI process is to determine who needs intervention through universal screening. Screening involves assessing all students with a measure of interest to determine whether additional support is needed to reach proficiency in that skill. There is considerable attention within the RTI research literature paid to screening students, which suggests that the measures used to do so should assess behaviors closely related to academic problems, should predict future academic outcomes (Jenkins, Hudson, & Johnson, 2007), and should align with the school's curriculum and instruction (Ikeda, Neessen, & Witt, 2008). Moreover, measures used to identify students who need additional support require a moderate level of psychometric adequacy, but there is room for some limited amount of error (Salvia, Ysseldyke, & Bolt, 2013).

The first decision within an RTI process is to determine who needs intervention through universal screening.

The National Center on Response to Intervention rated various tools used to identify who needs additional support for academic problems and provides their ratings at *www.rti4success.org/screeningTools*. The criteria with which each tool is rated include evidence for reliability and validity, adequate norms, and the diagnostic accuracy of the tool. The measures listed in Table 2.1 were rated as demonstrating convincing evidence in all areas. There currently is not widely available rating of behavioral screening tools, but the recently funded National Center for Intensive Interventions (*www.intensiveintervention.org*) will soon provide a vetted rating of behavior measures that have potential as screening tools.

Precision

Data used for screening purposes should give global estimates of the skill. In other words, screening tools should assess reading, math, writing, overall adaptive behavioral functioning, and so forth. This concept is called a general outcome measure (GOM). GOMs are tools that can allow for statements about global estimates of key skills. The classic educational example of a GOM is CBM of reading that provides an efficient method to estimate a child's oral reading fluency (ORF). ORF provides a reliable and valuable estimate of reading, which is foundation skill critical to school success. Many assessment tools give estimates of specific skills (e.g., decoding skills, single-digit multiplication, time on task), but relying on those specific skill measures might result in misidentifying too many students because they may

TABLE 2.1. Assessment Tools Rated by the National Center on Response to Intervention as Demonstrating Convincing Evidence for All Areas

	Screening	Monitoring progress—general outcome	Monitoring progress—skill
Reading	• Edcheckup Oral Reading Fluency • Predictive Assessment of Reading • Star Reading	• Aimsweb Oral Reading Fluency • Aimsweb Letter-Naming Fluency • Aimsweb Letter-Sound Fluency • Aimsweb Nonsense Word Fluency • Aimsweb Phoneme Segmentation Fluency • Star Early Literacy • Star Reading	
Math	• Star Math	• Star Math	• Accelerated Math • Math Facts in a Flash

be strong in that one aspect of learning or behavior measured by the tool, but may lack skill in other important areas.

Frequency

Annual assessments do not provide information that is relevant to intervention efforts primarily because of the infrequency with which those data are collected (Shepard, 2000). However, screening data are only used for low-level decisions (e.g., who needs additional assistance), and for this task periodic assessment data may be quite helpful. Thus RTI models should contain periodic assessments, often referred to as benchmark assessments or interim assessments, in which general outcome data are collected for every child three to five times each year.

Sensitivity to Change

The primary purpose of screening measures is to identify students who need additional support. However, school personnel can use screening data to measure overall program effectiveness and to estimate student growth within one school year. Thus there needs to be some sensitivity to change, but the measures occur as infrequently as once every 16 weeks. As a result, the measures used to identify students do not need to be overly sensitive to change. More important, the screening measures need to adequately differentiate among students. Sensitivity within a screening framework is estimated by how well the data dif-

ferentiate who will perform adequately on some future gold-standard criterion. More specifically, sensitivity is the how accurately a screening tool identifies students who will not perform well on a criterion measure, which for academic problems is frequently the state accountability test.

Identify Interventions

After students are identified as needing additional support, data are needed to help determine what intervention would most likely benefit each student because interventions that are closely matched to student skill result in improved learning and behavioral outcomes (Burns, 2007; Daly, Martens, Kilmer & Massie, 1996; Shapiro & Ager, 1992; Treptow, Burns, & McComas, 2007) and can have differential effects for individual students. As stated above, measures of cognitive processing, commonly referred to as an aptitude by treatment interaction, do not adequately inform the intervention selection process. However, a "skill by treatment interaction" (Burns, Codding, Boice, & Lukito, 2010), in which specific skills are measured to determine appropriate interventions, results in larger student effects. This is discussed more thoroughly in Chapter 3, and was extensively covered in Volume 1. Here we only discuss the type of measures that are appropriate for this process. Because these data are used for generating hypotheses, they would need to meet only a moderate level of psychometric adequacy. The data should probably be more reliable than those derived from screening measures (~0.70; Salvia et al., 2013), but a lower standard for diagnostic accuracy would be required. This lower level of reliability is acceptable, as any decisions will be subsequently tested.

Precision

Precision is a key component of measures used to identify interventions. It does not matter how well a child reads, for example; it matters how well he or she decodes *r*-controlled vowels, diagraphs, or diphthongs. In other words, reading fluency data are commonly collected in schools and can provide information about whether a child is struggling with reading fluency, but those data to not describe which specific skill deficits contribute to poor reading fluency or how to guide intervention. Thus intervention design decisions are heavily influenced by subskill mastery measures (SMMs), which assesses small domains of learning based on predetermined criteria for mastery (Fuchs & Deno, 1991). If a child struggles with reading, assessments of phonemic awareness and phonetic skills should be conducted to determine the appropriate intervention (Burns & Gibbons, 2012). Specific measures needed for mathematics skills could include knowledge of numeracy, basic fact fluency, and sign-to-operation correspondence (Fuchs et al., 2003). RTI has led to a resurgence in SMM with positive outcomes (Hosp & Ardoin, 2008; Ketterlin-Geller & Yovanoff, 2009; VanDerHeyden & Burns, 2009). Finally, in relation to behavior, focus is placed on understanding the specific function of the behavior (Iwata, Dorsey, Slifer, Baumann, & Richman, 1982; Mace, Yankanich, & West, 1988).

Frequency

Although research has consistently demonstrated that matching interventions to student skill leads to success, intervention design is at best a hypothesis-generating process. Thus hypotheses need to be developed and changed at a fairly rapid pace, and frequent skill measurement is needed. Data are not collected a particular interval, but are collected when changes in intervention are needed. For example, we may collect data regarding a student's math skills to suggest an intervention, attempt the intervention for 3 to 4 weeks, and then determine that modifications to the intervention protocol are needed. We would then collect additional data to further analyze the problem. We may decide that intervention modifications are needed after 2, 6, 8, or 10 weeks and our data collection system must allow that pattern.

Sensitivity to Change

Screening data can be used to monitor growth over the course of a year, and data used to monitor progress should show growth on a weekly basis, but data used to identify interventions must show immediate changes in behavior. Brief experimental analysis (BEA) is used to systematically and rapidly test the effects of different interventions and intervention components using a within-participant approach (Daly, Witt, Martens, & Dool, 1997). Using BEA, a few selected interventions are tried briefly (e.g., one to three sessions) and then evaluated to see which is most successful. This process allows for an intervention test drive both to see which intervention produces the most desired effect and to collect feedback on intervention fit from the teacher. Several studies have examined the effectiveness of BEA for identifying effective individualized interventions for improving academic skills and behavioral outcomes (Bonfiglio, Daly, Martens, Lin, & Corsaut, 2004; Burns & Wagner, 2008; Noell, Freeland, Witt, & Gansle, 2001), all of which relied on extremely sensitive data that could result in 80% increases in outcome data over the course of one intervention session (Burns & Wagner, 2008). This topic also links to the next point of determining intervention effectiveness, which is essentially an extended analysis of the intervention selected using this stage of assessment.

Determine Intervention Effectiveness

The final stage of any intervention model is to determine the effectiveness of the intervention, which is consistent with a summative evaluation paradigm (Bloom et al., 1971). However, data collected after or during intervention to determine effectiveness can be used for formative purposes by making changes to interventions based on a lack of student growth, and combining effectiveness data with screening and intervention design completes a comprehensive formative evaluation model. Although various measures can inform different aspects of an assessment to intervention model, CBM seems ideally suited to monitor progress for academic issues and direct behavior ratings (DBRs) for behavioral concerns; both of these are discussed in subsequent chapters.

There is considerable research examining the psychometric adequacy of CBM and DBR data when monitoring progress (Ardoin & Christ, 2009; Chafouleas, Sanetti, Kilgus, & Maggin, 2012; Christ, 2006; Riley-Tillman, Chafouleas, Sassu, Chanese, & Glazer, 2008; Riley-Tillman, Methe, & Weegar, 2009; Yeo, Kim, Branum-Martin, Wayman, & Espin, 2011). The reliability of the data depends on the decision being made. If the monitoring data are being used to modify an intervention, then a lower standard (e.g., 0.70 or 0.80) is needed than if the data are being used to make entitlement decisions. Although CBM data are generally sufficiently reliable for most decisions, there are characteristics of the progress-monitoring system that need to be in place for reliable decisions to be made (e.g., at least 8–10 data points; Christ, 2006). The National Center on Response to Intervention rates tools appropriate for monitoring progress when focusing on general outcomes (*www.rti4success.org/progressMonitoringTools*) or specific skills (*www.rti4success.org/progressMonitoringMasteryTools*). Several monitoring tools are rated on reliability and validity of the scores, but also of the slope of growth and rates of improvement. Several measures meet all criteria for effective progress monitoring, most of which are commercially available CBM packages, and are listed in Table 2.1. As with screening measures, there currently no widely available rating of behavioral progress-monitoring tools. Luckily, the recently funded National Center for Intensive Intervention (*www.intensiveintervention.org*) will soon provide a vetted rating of behavior measures that have evidence to support their use for behavioral progress monitoring.

Precision

There is a need for precision and generality when monitoring progress. Practitioners should use both GOMs and SMMs when evaluating the effectiveness of an intervention. GOMs provide information about progress in the global skills (e.g., reading), but SMMs demonstrate progress or lack thereof in the skill that the intervention is targeting.

An example of the need for both measures comes from the experience of the second author (M.K.B.) while working in a K–2 elementary school. Each month the school team met to examine intervention effectiveness data and to problem-solve any difficulties. One of the special education teachers presented data for all of her students that consisted of oral reading fluency CBMs and were presented in individual graphs. Unfortunately, the data over the previous 4 weeks suggested a flat rate of growth, and the frustrated teacher stated, "I'm doing a decoding intervention, and by the way, Dr. Burns, it is the decoding intervention that you recommended, and my students aren't doing well. What do I do?" We examined the data and concluded that a decoding intervention was probably appropriate based on our intervention-design data, but that not enough time had passed to see an effect in this general measure. We recommended that the teacher still collect the ORF data because they are good measures of overall reading proficiency, but that she also collect data regarding the students' decoding skills. She then started collecting nonsense word-fluency data on a weekly basis in addition to the ORF data, and an immediate growth in skill was noted.

As can be seen from the example above, a skill measure was able to show growth before the general measure did. Thus relying only on GOMs might have resulted in prematurely

abandoning an intervention. Alternatively, relying only on SMM to monitor intervention effectiveness would not indicate increases, or lack thereof, in global skills and may result in maintaining an intervention too long when there was a need to further accelerate growth.

Frequency

Data used to monitor progress should be collected frequently. As stated above, research by Christ (2006) found that a minimum of approximately eight data points is needed to make reliable decisions. Thus if data are collected only once per week or less, much more time will be needed to obtain sufficient data to make a reliable decision. Data are often collected weekly, and once every other week should be considered a minimum for effective data-based decision making. If a team wishes to discuss the effect on an intervention after several weeks, an appropriate data collection schedule should be designed. This is one reason why it is critical that progress monitoring measures are designed so that they can be efficiently collected. For example, a CBM probe can only take a few minutes to administer, while DBR probes can take less than a minute for each data point. Different types of data within one progress monitoring system could be collected at different intervals. For example, SMM data could be collected weekly, but well-constructed GOMs such as ORF could be collected every other week, and even more global measures (e.g., an individually administered measure of reading comprehension) could be collected once each month.

Sensitivity to Change

Because the goal of progress monitoring data is to document changes in a target behavior, data used for this purpose must be sensitive to change. Traditional assessment practices are often criticized for a lack of instructional utility because they lack overlap between assessment and curriculum, and are insensitive to changes in behavior. Student performance on norm-referenced tests is interpreted in comparison to a norm group, which makes it difficult to obtain changes in scores between test administrations.

Measures designed for progress monitoring purposes such as CBM and DBR are generally more sensitive to change than most standardized measures used in schools. SMM data are often more sensitive to change than GOMs, but the latter should also be sufficiently sensitive to model short-term growth in the global skill, which again supports the argument for collecting both types of data within a progress monitoring system.

Many measures used to monitor progress are timed (e.g., ORF), and we frequently field questions from classroom teachers about the need to time these measures. Yes, there are legitimate concerns about timing assessments for some students, but there are several distinct advantages. First, timing the measure increases its standardization, which is important if the data are to be used for important decisions. Second, timing the measure makes it much more sensitive to change. Consider two third-grade students who are struggling learning their single-digit multiplication facts. Both students are assessed with the facts for the 3's, 4's, and 6's and are allowed 2 seconds to respond to each one. If the student responds correctly within 2 seconds, then the fact is counted as correct, but incorrect responses or

responses given after 2 seconds are counted as errors. There are 10 facts each, for a total of 30. Assume one student has no trouble and correctly states the answer for all 30 within 2 seconds each and completes the task with 100% accuracy in about 28 seconds. The other student also can state the answer for each one, but he must think about it much more thoroughly. The second student correctly answers all 30, but requires 54 seconds to do so. Thus both students have a score of 30, or 100% correct. However, do these two students have the same level of mastery of the problems? The answer is no, and timing the assessment is the only way to determine the difference between the two sets of skills. Moreover, if an intervention helps a student go from completing all 30 facts in 54 seconds to completing all 30 facts in 28 seconds, then those data suggest an effective intervention, which would not be seen in comparing 100% to 100%.

Determine Whether Special Education Is Warranted

Although the focus of multi-tiered system of support is and should be on using data to enhance student learning (Burns & VanDerHeyden, 2006), the application of the construct to an RTI model came from special education identification. IDEIA 2004 allowed schools to use a process that determines whether the child responds to *scientific, research-based interventions* as a part of the evaluation procedures for learning disability (LD) eligibility determination. Data used for LD identification must be held to the highest standards for reliability (~ 0.90; Salvia et al., 2013) and validity. A review of research, policy documents, and ethical guidelines suggested that RTI-based assessment practices, when carefully implemented, have the potential to be multifaceted, fair, valid, and useful (Burns, Jacob, & Wagner, 2008). However, there were legitimate threats to acceptable RTI-based assessment practices including poor treatment fidelity; a lack of research-based interventions appropriate for diverse ethnic groups, older students, and students with limited English proficiency; and inconsistent definitions of nonresponse to intervention and when that would warrant formal referral for evaluation of special education eligibility (Burns, Jacobs, & Wagner, 2008).

The issues of precision, frequency, and sensitivity are less relevant for this decision than for the other three, partly because the other three decisions happen within the RTI framework and this final one (LD identification) is the result of the process. Thus the precision, frequency, and sensitivity are determined within the different decisions made during the process, and LD identification is not the outcome of the model. In other words, some practitioners in the schools where we work express concern that identifying students as LD with an RTI process will slow down the identification process because they need to attempt a Tier 2 intervention for 8 weeks or so, then a Tier 3 intervention for another 8 to 10 weeks. Unfortunately, this proposed progression reflects a misunderstanding of RTI because LD identification does not happen by starting an RTI process when a difficulty is suspected; it happens by examining the data that already exist. Therefore, the technical adequacy of the model and

LD identification does not happen by starting an RTI process when a difficulty is suspected. It happens by examining the data that already exist.

the data collection procedures should be evaluated within the context of the specific decisions within the model.

IDEIA requires "a full and individual initial evaluation" prior to providing special education services (Public Law 108-466 § 614 [a][1][A]), which could include health, vision, hearing, social and emotional status, general intelligence, academic performance, communicative status, and motor abilities, if appropriate. Collecting RTI data is not in and of itself a comprehensive evaluation, but additional data are collected only when appropriate. According to the 2006 *Federal Register*, U.S. Department of Education personnel stated in the comments section that

> the Department does not believe that an assessment of psychological or cognitive processing should be required in determining whether a child has an SLD [specific learning disability]. There is no current evidence that such assessments are necessary or sufficient for identifying SLD. Further, in many cases, these assessments have not been used to make appropriate intervention decisions. (p. 46651)

Thus comprehensive evaluations for LD identification using RTI data may be determined by appropriate precision, frequency, and sensitivity within the model and not whether cognitive processes were measured. We discuss LD identification extensively in Chapter 9 and discuss evidence-based interventions in Volume 1.

TREATMENT FIDELITY

Valid decision making is the primary criterion by which all assessment data are judged (Messick, 1995). Treatment fidelity has been repeatedly identified as the greatest threat to valid decisions within an RTI model (Burns, Jacobs, & Wagner, 2008; Noell & Gansle, 2006). Consistent and correct implementation of interventions is necessary to assure substantive intervention plans (Noell & Gansle, 2006). The relationship between treatment plan implementation and outcome is complex (Noell, 2008), but generally speaking, treatments become increasingly likely to lose effectiveness or fail entirely as implementation integrity decreases (Gansle & McMahon, 1997; Noell, Duhon, Gatti, & Connell, 2002; Vollmer, Roane, Ringdahl, & Marcus, 1999). It is also worth acknowledging that some omissions in treatment implementation appear to be less critical and that some imperfections in implementation are likely to have little practical consequence (Noell & Gansle, 2006).

Intervention fidelity is a multifaceted construct that should be assessed with multiple sources of data (Sanetti & Kratochwill, 2009). Most assessments of treatment fidelity consist of direct observations of interventions while following intervention protocols to determine whether the steps of the intervention are in place (Sanetti et al., 2011), but that treats fidelity like a unidimensional construct. Instead, a four-pronged approached is recommended that includes examining permanent products, directly observing the intervention, self-monitoring and self-reporting, and using manualized treatments and intervention scripts (Sanetti & Kratochwill, 2008). Using manualized interventions that include intervention

scripts could lay the foundation for treatment fidelity assessments because the scripts could be used to judge the permanent products, to observe the intervention, and to complete self-reports. The intervention protocols included in Volume 1 were designed to assist as scripts, or as the basis for scripts to be created, from which implementation integrity could be assessed.

Permanent Product

Perhaps the most basic component of a treatment fidelity plan is to examine permanent products. Just about any intervention that occurs in K–12 schools results in something being created. For example, reading interventions may use student workbooks and assessments, initial placement worksheets, student books, and so forth. Behavioral interventions include products such as completed behavioral reports, markings on a board, and completed token economy sheets. Moreover, computer-based interventions are ideal for examining permanent products because they often include daily or weekly point sheets or can easily record the number of activities completed. Although the presence of permanent products does not ensure that an intervention was implemented with fidelity, the absence of the products suggests an intervention fidelity issue.

Direct Observation

Intervention scripts can be easily converted into implementation checklists with which practitioners can observe interventions. Although there is no research-based minimum, integrity checks generally involve observing 20–25% of the intervention sessions. This may be an unrealistic goal in an applied setting, and using a multidimensional approach reduces the required frequency with which the interventions must be observed, but 10% seems to be a reasonable minimum expectation. If these checks are random and unscheduled, they are logically more likely to accurately estimate typical levels of implementation integrity.

Perhaps it is more important to ask how much integrity is needed, rather than how often interventions should be observed. It seems that 90% integrity would assure effective implementation, but that is not always the case. For example, if a teacher were to implement a behavior plan that involved reinforcing alternative behavior, which was observed with a 10-item checklist for which providing the reinforcement was one item, then 90% integrity would not be sufficient if the one item that was not observed was the actual provision of the reinforcement. Therefore, intervention teams are encouraged to discuss (1) how often the observation should occur, (2) what is the minimum level of implementation integrity that would be judged as sufficient, and (3) what items are most critical to success. One of the components of the intervention briefs is Volume 1 is related to this issues. Those briefs presented "critical components" for each intervention. In terms of integrity checks, those critical components must be present for integrity to be considered sufficient. Ideally, all interventions will be presented in this manner over time to help educational professionals know what elements are essential and what elements can be altered.

Self-Report

Previous research found that teachers can accurately self-report treatment fidelity (Sanetti & Kratochwill, 2009). Thus intervention protocols can also be used as a self-report implementation checklist in which the teacher reports whether each aspect of the intervention plan was implemented. Once again, there is no hard-and-fast rule about how often self-report should occur or how much integrity is needed. Self-report data can be easy to collect, but it adds another requirement to the implementer. Thus self-report data may be collected only periodically, such as weekly, but brief self-reports could be collected daily or at every intervention session.

Implementation Integrity System

Implementation integrity is important, but does not happen by accident, and neither does the assessment thereof. School-based teams must carefully design a process for assessing integrity while initially developing the intervention and intervention plan. Sanetti and Kratochwill (2009) described a three-stage process in which teams first define the intervention and the necessary steps within it, then collaboratively plan the integrity assessment plan, and finally create a self-assessment from the information included in the previous phases. This process should be implemented at various levels of the RTI process.

Grade-Level Teams

In our experience, many schools start the RTI implementation process by starting a problem-solving team (PST). We can promise that if your first step in implementing an RTI framework is to start a PST, then your model will be doomed to failure. PSTs are a powerful aspect of an RTI model, but have two shortcomings. First, as we discuss below, most PSTs do not solve problems—they admire them. Second, starting a PST as the first step likely ignores the most important point. On average, 20% of students require more assistance than they receive in a typical general education curriculum (Burns, Appleton, & Stehouwer, 2005). Consider an elementary school of 500 students. If a PST meets to discuss any students experiencing a problem, then they will meet to talk about 100 students (20% of 500 = 100), which is far too many to conduct an in-depth analysis and to implement individualized interventions. Instead, schools are wise to start by examining the quality of the core curriculum. Thus it is not the PST that drives an RTI process, it is the grade-level team (GLT).

We discuss effective GLTs in Chapter 6. For this conversation we will assume GLTs meet to discuss Tier 1 difficulties, to identify who needs Tier 2 interventions, and to evaluate the progress of students receiving Tier 2 and Tier 3 interventions. See below for discussion about schools without GLT. We suggest that GLTs meet on a weekly basis to discuss Tier 1 and Tier 2 interventions and that they identify two types of evidence for the interventions. First, the outcome data should be discussed to determine how they will assess progress and student learning. Second, they should identify process goals in order to monitor the prog-

ress with which the instructional activities or interventions are implemented, and at that monthly meeting they develop an implementation integrity assessment plan.

Table 2.2 lists the phases of the intervention assessment plan and how a GLT could implement it. Permanent products for instructional activities can consist of items like student workbooks, but could also include lesson plans and student assessments. We also frequently record instructional lessons in order to view and discuss the lesson at future GLT meetings.

The GLT team also evaluates student progress and implementation integrity associated with Tier 2 interventions. Most commercially prepared interventions appropriate for Tier 2 include some sort of implementation checklist that the team can adopt and convert into a self-report assessment. Moreover, we (Burns, Deno, & Jimerson, 2007) examined research to determine the components of an effective Tier 2 intervention and created a generic obser-

TABLE 2.2. Example of an Intervention Integrity Assessment Plan for GLTs at Tiers 1 and 2, and PSTs at Tier 3

	Phase 1: Define intervention	Phase 2: Develop a plan	Phase 3: Self-assessment plan
GLT instructional lesson (Tier 1)	Teachers select the activity and determine essential components.	Teachers collaboratively determine permanent products (e.g., lesson plans) and find implementation checklists (e.g., curriculum guide or protocols).	Teachers determine what will be discussed at next GLT meeting.
GLT small-group intervention (Tier 2)	Teachers select the intervention and determine essential components.	Teachers collaboratively identify permanent products from the intervention (e.g., student workbook), and create or find an intervention protocol.	Teachers convert intervention protocol into a self-report and determine how often each aspect of the integrity plan will be implemented. The data are shared at each subsequent GLT meeting.
PST interventions and PST process (Tier 3)	PST identifies intervention and essential attributes.	PST ends every meeting by deciding what will serve as a permanent product, of what the intervention protocol will consist, and how the interventionist will self-report. PST also identifies a checklist for the PST process and selects the essential items from it.	Interventionist completes the self-report and all data are reported back to the PST at the follow-up meeting. The PST ends each meeting by deciding whether they implemented the four or five items from the PST process implementation checklist that they judged to be most important. The PST decides whether they implement all of the items from the PST process implementation checklist on a periodic basis.

vation checklist that is available in Form 2.1, at the end of this chapter. Team members could also identify permanent products within the intervention protocol at the monthly GLT meeting. If the intervention is not commercially prepared, then the GLTs would have to identify the essential components of the intervention, develop an intervention protocol, and convert that protocol to a self-report.

It is important to address schools without GLTs. For example, small rural schools may have one teacher per grade. Another common example is at the middle- and high school level, where there may be department teams rather than GLTs. In such situations some group should commonly meet to fill the GLT role. It is impossible to outline a model that will work for every school, but we do feel that once the role of GLT is understood, a good principal can apply those responsibilities to a logical team of educational professionals.

Problem-Solving Team

The RTI process is driven by the GLT, but a high-functioning PST is critically important for developing interventions within Tier 3 after reviewing relevant data from Tiers 1 and 2. Thus, the PST should (1) identify the intervention, (2) decide how progress will be monitored, (3) determine what permanent products will be created, (4) select or create an intervention protocol, (5) convert the intervention protocol to a self-report, (6) determine how often each integrity assessment will occur, (7) write the intervention integrity assessment plan into the intervention plan, and (8) examine the integrity plan data at a follow-up consultation and at the follow-up meeting.

In addition to intervention integrity, the integrity with which the problem-solving process was implemented should be examined. Burns, Peters, and Noell (2008) created a 20-item checklist and used it to provide performance feedback to the team. Simply providing performance feedback increased the integrity with which the items were implemented. However, some important items were not implemented (e.g., develops an intervention plan with the teacher). Thus the PST should identify four or five items that take priority. They could be items that the team judges to be most important, or they could be items that the team judges to be weaknesses for that particular team. Then the PST would judge whether they implemented those four or five items at the end of every meeting, and would periodically complete the entire checklist (e.g., once a month, once a quarter, once a semester). The self-report implementation data should be stored somewhere as documentation that the PST process occurred.

CONCLUSION

RTI can be conceptualized as the use of assessment data to systematically and efficiently allocate resources for the purpose of improving learning for all students (Burns & VanDerHeyden, 2006). Thus those of us who are passionate about assessment see this as an opportunity to finally use school-based assessment data to their fullest potential. We know what data are needed in order to conduct assessment *for* learning (Stiggins, 2005), and we know

how those assessments should be conducted and the data used within a multi-tiered system of support. Data should be used to identify who needs additional intervention, what intervention is needed, to determine whether the intervention is effective, and to determine whether special education services are needed. However, data for these important decisions will not lead to valid conclusions within an RTI framework unless the interventions are implemented with fidelity, which is assessed with the combination of permanent products, direct observation, and self-report.

Data should be used to identify who needs additional intervention and what intervention is needed. Then, data will determine if the intervention is effective and if special education services are needed.

Effective data collection efforts and treatment fidelity are keys to successful RTI implementation. In subsequent chapters we discuss specific data to examine at Tiers 1, 2, and 3 to answer the questions outlined in this chapter, and we discuss personnel who should examine the data and the process to do so. All of these factors combined lead to multi-tiered systems of support that address the learning needs of all students.

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

FORM 2.1

Generic Tier 2 Fidelity Observation Checklist

Item	Observed	
The Tier 2 intervention is:		
1. Implemented or supervised by a qualified teacher with reading expertise	Yes	No
2. Targeting one specific reading skill	Yes	No
3. Targeting a skill that is consistent with one of the five areas identified by the National Reading Panel	Yes	No
4. Implemented 3 to 5 times/week	Yes	No
5. Implemented in 20- to 30-minute sessions	Yes	No
6. Delivered in a small-group format	Yes	No
7. Occurring in addition to core reading instruction	Yes	No
8. Designed to last at least 8 weeks	Yes	No
9. Monitored with a rate-based measure and slope of student reading growth	Yes	No
10. Evidence based (at least a moderate effect size)	Yes	No