CHAPTER 3

CBA-ID as an Assessment Tool

Effective teaching is, at least in one sense, the melding of assessment and instruction. There is a clear and strong link between assessment and teaching to the point that monitoring student learning is a core proposition for effective teaching by the National Board for Professional Teaching Standards (*www.nbpts.org/five-core-propositions*), and assessment is a basic standard for special education teachers (Council for Exceptional Children Board of Directors, 2004). When we were both practicing school psychologists we often taught students new information as a form of assessment to the point where onlookers inquired about the interventions we were conducting with students who were referred for a special education disability identification evaluation. Our response was "How can you determine if a child has a learning disability unless you watch the child during learning or without seeing if learning can occur?"

If assessment is so important to instruction, then why do teachers so often cringe at the sound of the word? It is because the educational community has lost sight of the power of assessment and has focused on the consequences of judgments made with the data. It may also be because of an unfortunate emphasis placed on summative approaches, which focus on producing data that summarize prior learning over formative approaches, which focus on producing data to inform future instruction (Stiggins, 2005). Determining if your students have met specific or global proficiency standards certainly plays a role in designing instruction because educators need to know whether our students learned the material to determine if we should reteach it. However, summative approaches provide minimal information for actually designing instruction. The accountability movement has changed teachers' and parents' perceptions of the word *assessment*.

When we talk about assessment, we are essentially talking about formative evaluation. The term *formative evaluation* is often misused to mean assessment that occurs before learning occurs, or monitoring student progress during learning (Linn & Gronlund, 2000; Salvia, Ysseldyke, & Bolt, 2007). In fact, curriculum-based measurement (CBM) has become synonymous with formative evaluation (Deno, 2003; Silberglitt & Hintze, 2005) because

CBM is so closely linked to monitoring progress. However, formative evaluation is best conceptualized as using data to identify student needs and to plan instruction that will better meet those needs (William, 2006). Formative evaluation is an ongoing feedback loop that simultaneously measures current student functioning and dictates future instructional activities. Progress monitoring is only one purpose for which data are used in the instructional process. CBM produces longitudinal data that are an excellent indicator of the rate of learning, but it seems to have somewhat limited utility in identifying specific strengths and difficulties for individual students (Fuchs, Fuchs, Hosp, & Hamlett, 2003). Certainly we should be monitoring student progress in the name of formative evaluation, but if that is all we are doing, than we are just barely tapping into the potential of a powerful instructional tool.

Formative evaluation is best conceptualized as using data to identify student needs and to plan instruction that will better meet those needs.

CBA-ID AS A CORE COMPONENT OF FORMATIVE EVALUATION

Algozzine, Ysseldyke, and Elliott (1997) presented a model of effective instruction that included (1) planning instruction, (2) managing instruction, (3) delivering instruction, and (4) evaluating instruction. Formative evaluation should address all four phases, but much of the research attention has been on evaluating instruction. CBM seems ideally suited to evaluate the effectiveness of instruction and interventions, but CBA-ID seems better suited to plan, manage, and deliver instruction.

Planning Instruction

Setting goals is one of the basic tasks within planning instruction. Thus, educational professionals have developed several approaches to determining norm-referenced goals for CBM data. Normative goals are potentially useful for identifying students who need additional support (screening), but they provide little information to determine if a student has reached a level of proficiency. Does reading at the 50th percentile represent proficient reading? It may if the 50th percentile represents performance within the local norms of a high-achieving school or district, but in low-achieving schools or districts a student at the 50th percentile of the local norms may still be a very low reader. The point is that normative data alone do not adequately indicate successful skills.

A second component of planning instruction is deciding what to teach by "assessing skill levels to identify gaps between actual and expected level of performance" (Salvia et al., 2007). Although taking a fluency probe with CBM and comparing those data to local norms could accomplish this, this process does little to suggest *how* to close the gap between actual and expected performance. CBA-ID links data to intervention by suggesting the need for more or less challenging instructional material or by identifying specific items (such as unknown words in a reading curriculum), which can then be taught to the individual student. The instructional level of 93–97% known for reading is well researched

(Burns, 2007; Gickling & Armstrong, 1978; Shapiro, 1992; Shapiro & Ager, 1992; Treptow et al., 2007) and could serve as a potential criterion for planning what to teach. Students reading at a percentage of known material that fell below 93% could participate in efforts to increase the percentage of known words until the 93–97% known range is obtained, or students within the 93–97% range could participate in efforts to increase fluent reading. For example, a student in upper elementary could feasibly read 19 out of 20 words correctly (which is a low score for second grade and beyond) and still be within the instructional level (95%), but these data are much more useful than normative data alone. The normative data identify that the student is low in reading skills, but the CBA-ID data indicate that a proficiency-focused intervention is likely best because the student already reads accurately.

Finally, planning instruction involves pacing instruction appropriately. The longitudinal data of CBM are helpful for determining the student's pace of learning, but CBA-ID are more helpful for informing the teacher's pace of instruction. For example, ARs can be used to suggest the appropriate pace of instruction by determining how many unknown items can be taught before retroactive cognitive interference occurs. These data can be used to identify text levels that produce an appropriate number of unknown items so that learning can be maximized, but not frustrating.

Managing Instruction

Many of the tasks described as relevant to managing instruction involve various classroom management activities such as setting rules, teaching compliance, handling disruptions, and establishing a positive classroom environment. However, using time productively and maintaining academic focus were also emphasized. Teaching children at their instructional level increases task completion, task comprehension, and time on task (Gickling & Armstrong, 1978; Treptow et al., 2007), and exceeding a student's AR led to increased off-task behavior (Burns & Dean, 2005a). In fact, CBA-ID "is structured to help teachers plan instruction based on entry-level skills of students, thus maximizing on-task time during learning activities" (Gickling & Rosenfield, 1995, p. 588). There is certainly more to managing instruction and ensuring optimal academic learning time than using CBA-ID, but increasing time on task and decreasing behavioral difficulties are major components.

Delivering Instruction

Algozzine and colleagues (1997) also described showing enthusiasm, helping students value schoolwork, using rewards effectively, and modeling correct performance as part of the delivery of instruction. Activities more relevant to assessment practices, specifically CBA-ID, include assigning the appropriate amount of work, monitoring performance regularly, and providing opportunities for success while limiting opportunities for failure.

Standards for the assessment of reading and writing established by the International Reading Association (IRA) and the National Council of Teachers of English (NCTE; 1996) suggest that assessment data based on tasks that are either too easy or too difficult are not instructionally useful. Because CBA-ID assesses skills using tasks that match the individual student's skills, the data meet the IRA and NCTE standards and may be instruc-

tionally useful. In addition, most assessment models require some aspect of student failure (Hargis, 2005), but a basic goal of CBA-ID is obtaining high success rates for all students. Finally, teaching students at their individual instructional level increases student success and reduces the likelihood of student frustration (Gravios & Gickling, 2002). Thus, CBA-ID directly provides opportunities for success and seeks to limit student failure.

TYPE OF ASSESSMENT

CBA-ID can be useful to plan, manage, and deliver instruction, but has less utility to monitor progress. That is because CBA-ID measures specific skills rather than general outcomes. Within assessment, special education, and school psychology literature, there are generally two types of measures delineated: general outcome measures (GOMs) and subskill mastery measures (SMMs). A GOM is a standardized measure that assesses proficiency of global outcomes associated with an entire curriculum, and an SMM assesses smaller domains of learning based on predetermined criteria for mastery (Fuchs & Deno, 1991). The GOM and SMM data are each part of an integrated system of instructionally relevant data collection (Shapiro, 2011), and studies show positive outcomes when used together (Burns, 2002; Shapiro & Ager, 1992). We will discuss both below.

General Outcome Measures

GOMs are assessments of general outcomes and are often used to monitor progress, which involves frequently assessing children's academic development in order to make changes to instruction based on progress or a lack of progress (Speece, n.d.). The goal of GOM is to assess instructional effectiveness and quickly make changes as needed. Therefore, GOMs tend to be appropriate for and used as summative evaluations because the data are used to judge the effectiveness of instruction and may suggest a need for change. However, GOM data do not suggest what change is needed, only that one should occur. In fact, the Bloom, Hastings, and Madaus (1971) definition specifically lists "evaluation of progress" (p. 117) as an example of summative evaluation. GOM data can be critically important to the instructional process, but do not represent formative evaluation in and of themselves.

GOM data become more formative in nature when they are used to establish goals, which are important aspects of instructional planning (Algozzine et al., 1997), but summative evaluation samples learning tasks and formative evaluation examines all important aspects of the specific leaning unit (Bloom et al., 1971). Thus, formative evaluation probably cannot rely entirely on GOM data.

Subskill Mastery Measures

SMMs are more closely aligned with formative evaluation than GOM data because they are used to directly assess the learning unit to identify student strengths and needs before instruction occurs. For example, using CBA-ID to examine the percentage of words within

an upcoming reading task can determine if the task will present an appropriate challenge, or if it will be too easy or too difficult. CBA-ID is an SMM because it focuses on specific skills such as reading a particular passage or book, completing a specific math objective (e.g., single-digit multiplication), or focuses on one particular aspect of writing. Moreover, data from CBA-ID are compared to a mastery criterion (i.e., the instructional or independent levels) rather than to a norm group. There are no percentile ranks for CBA-ID.

SMMs are closely aligned with formative evaluation.

The focus in special education and school psychology research and practice has been on GOM because school psychologists were frequently involved in monitoring student progress. Moreover, the psychometric properties of GOM data tend to be stronger than SMM, or at least better established. However, as discussed below, there is considerable support for the reliability of CBA-ID data and the validity of the resulting decisions. Thus, teachers, school psychologists, and interventionists interested in conducting formative evaluation could consider CBA-ID an SMM that fits well into their assessment arsenal.

RELIABILITY AND VALIDITY OF CBA-ID

CBA-ID provides data that can be used to make formative evaluation decisions and teachers can use those data throughout the instructional process. However, standards regarding the use of educational assessments published by the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education (1999) state that the data derived from measures used in education should be reliable and should result in valid decisions. Teachers may not fully appreciate the importance of test reliability, but it is a concept with very real implications for classroom decisions. Every test has some level of error in it. Think back to a time in college or high school in which you did not prepare for a test quite like you should have. We are sure that even the most dedicated scholar occasionally did not adequately prepare for a PSYCH 100 exam or let that NAT SCI 201 exam sneak up on him or her. Now, of the times that you did not study as well as you should have, did you ever leave the exam session unsure how you did, but then received a pleasant surprise when the tests were handed back when you discovered that you actually did quite well? A more memorable alternative might be the time that you really prepared well for an exam, left the exam session confident that you aced the test, only to discover that the professor did not agree with your summation, which was reflected in a less-than-expected/desired grade.

In the two scenarios described above, did you feel that the test accurately represented your true knowledge? Every person has a true score that represents their true knowledge, skill, or aptitude. However, the observed score may not always exactly match a person's true score. The difference between the true score and the observed score is test error. Every test has some level of error. The question is how much error is acceptable? If you are using the test to determine whether you should reteach something, a relatively higher level of error would probably be fine because the consequence for getting it wrong would not be great.

However, if you used the data to decide if the child should be retained in the grade, be admitted into a graduate program, receive a monetary scholarship that could make college accessible, or be placed into special education, then the consequences for getting it wrong would be substantial and a much smaller amount of error would be acceptable. Reliability is an estimate of how free a test score is of error. In other words, a reliability of .90 suggests that 10% of the score is due to error. A reliability of .75 would have 25% error. Imagine telling parents that the score on a test suggests that their child has a reading disability when that score has 25–30% error!

Every test has some level of error. The question is how much error is acceptable?

Data are considered reliable if they are consistent across time, forms, and scorers. Establishing the validity of decisions made with the data is a more complex process. The reliability and validity evidence for data derived from CBA-ID are discussed below.

Reliability

Reading

Burns and colleagues (2000) examined reliability estimates for assessing reading skills with CBA-ID using 93 general education students from second, third, and fourth grades. Results included interscorer reliability coefficients that ranged from .89 to .99, internal consistency coefficients of .87 to .96, alternate form-reliability estimates from .80 to .86, and test–retest coefficients that ranged from .82 to .96 for a 2-week test–retest interval. These data suggested that the percentage of known words within a reading task could be reliably measured across time, forms, and scorers. However, these data also examined reliability a second way in that the test–retest reliability estimates were computed by converting the data from a raw percentage to a category of frustration (less than 93%), instructional (93–97%), and independent (98–100%). The reliability of the category was then calculated by correlating the categorical score with a Tau coefficient, which resulted in coefficients above .80. Salvia, Ysseldyke, and Bolt (2010) suggested that coefficients of .80 or higher are needed for screening decisions about individual students and .90 or higher for important decisions (e.g., special education eligibility) about individual students. Thus, these data suggest that CBA-ID data from reading were sufficiently reliable for instructional decisions.

CBA-ID data from reading were sufficiently reliable for instructional decisions.

Writing

The focus of the book to this point has been reading because most of the research around CBA-ID was in reading. However, there are CBA-ID procedures for writing and math as well. Early writing skills can be measured with different prompts depending on the student's skill. Picture–word prompts provide a word with a picture above it, and students write a sentence using the word provided. Sentence copying prompts provide simple sen-

tences that students copy on lined paper. Both prompts require students to write for 3 minutes. Student responses are scored by counting the number of words written (WW), words spelled correctly (WSC), and correct word sequences (CWS). The assessment procedures are explained in detail in Chapter 6. Reliability estimates for these data all met or exceeded .70 across 2 weeks (Parker et al., 2011).

Math

CBA-ID for math is conducted by measuring specific skills within a curriculum. For example, a practitioner would assess a student using a timed probe of single-digit multiplication problems, but then would compute the number of digits correct per minute and compare that with an instructional-level criterion. Math assessment procedures are discussed in Chapter 5. Previous research among students in second through fifth grade found reliability coefficients of .64 for second and third grade and .85 for fourth and fifth grade (Burns et al., 2006). Five of the seven coefficients met or exceeded .70, and three exceeded .80. Moreover, the categorical data of frustration, instructional, or independent levels correlated across time with a coefficient of .42 for second and third graders, and .71 for fourth and fifth graders. Thus, data obtained from CBA-ID for math were sufficiently reliable for instructional decisions.

Acquisition Rate

The second dimension of CBA-ID, measuring ARs, was examined by Burns (2001) through estimates of delayed-alternate form reliability using sight-word recognition as the academic task. A total of 91 students from first, third, and fifth grades were taught unknown sight words until interference occurred, with the number of words learned and retained being recorded as the AR. The process was repeated 2 weeks later using different unknown sight words. Reliability estimates for ARs were .76 for first-grade students, .91 for third-grade students, and .91 for students in the fifth grade. The delayed-alternate form reliability coefficient for the total sample was .93. These coefficients suggested adequate reliability for instructional decision making.

Validity

Establishing the validity of decisions made with assessment data is more complicated than reliability. Content relevance is often considered a critical component of valid academic assessments (Messick, 1995), and is conceptualized as the extent to which the domain being measured is represented. One of the basic tenets of CBA-ID is that the assessment is ensured to match the curriculum because curricular content and objectives form the materials for the assessment. Unlike other curriculum-based

Unlike other curriculum-based approaches (e.g., CBM), CBA-ID does not utilize alternative curricula or standardized probes from a pool but instead uses the same curricular material for assessment and subsequent intervention.

approaches (e.g., CBM), CBA-ID does not utilize alternative curricula or standardized probes from a pool but instead uses the same curricular material for assessment and subsequent intervention (Tucker, 1985).

Criterion-related validity is the extent to which data from one measure are related to data from an existing measure of the same or similar construct. Data obtained from math CBA-ID correlated with a standardized measure of math at .55, and the categorical data (frustration, instructional, or independent levels) correlated with the math measure at a coefficient of .14–.52 (Burns et al., 2006). Writing CBA-ID data correlated with a standardized measure of writing at coefficients that ranged from .26 to .52 (four out of six coefficients exceeded .40) for the raw data and .21–.50 (four out of six coefficients exceeded .40) for the categorical data (Parker et al., 2011). Finally, data obtained from measuring ARs with CBA-ID correlated with a standardized measure of memory at .70 with third- and fourth-grade students.

Construct validity is the extent to which a test or assessment procedure measures the theoretical trait or characteristic it purports to measure (Salvia et al., 2007). CBA-ID purports to measure the instructional level, which was defined by Gravois and Gickling (2002) as "a comfort zone created when the student has sufficient prior knowledge and skill to successfully interact with the task and still learn new information" (p. 888). Research has consistently demonstrated that using material that aligned to an instructional level for reading or math resulted in increased student learning (Burns, 2007; Treptow et al., 2007; VanDerHeyden & Burns, 2005b).

Research and theory regarding acqusition rates seems to be consistent with previous memory research. Several scholars have examined the limits of human learning (e.g., Fry & Hale, 1996; Gathercole & Baddeley, 1993; Miller, 1956). Brainerd and Reyna (1995) identified individual differences among students in their ability to acquire and retain new information, which suggest an individual capacity that might be affected by the content of the information (Scweickert & Boruff, 1986) or individual experience with the topic or data (Rabinowitz et al., 1994). Gregory (2000) argued that validity for assessment data could be suggested by evidence for consistency with expected developmental changes. ARs have been studied by cognitive researchers (Fry & Hale, 1996; Gathercole & Baddeley, 1993; Miller & Vernon, 1996), who have found a developmental effect on working memory with older children being capable of acquiring and retaining more information as compared with younger children. Burns (2004a) found a similar developmental trend for ARs of sight words, as measured with CBA-ID, but suggested that age no longer adequately predicted ARs after third grade. This latter finding was consistent with Gathercole and Baddeley (1993), who found that active rehearsal and consistent retention rates occurred between the ages of 6 and 8 years, which led to consistent individual differences in retention among students after the third grade. In other words, developmental effects accounted for variance among students until ages 6–8 years, afterward individual differences in rehearsal strategies were consistently used by students and accounted for differences in memory capabilities.

Of course, the ultimate test of how valid decisions are in education is how well using the data improve student learning (Kane, 2001). Modifying instruction based on CBA-ID data among students with a learning disability resulted in reading growth rates that exceeded

students without disabilities for 66% of the students, and all students saw increases in their growth rates (Burns, 2007). Similar increases have also been noted for math (Burns, 2002; VanDerHeyden & Burns, 2005b), and exceeding a student's AR resulted in an immediate and dramatic increase in off-task behavior (Burns & Dean, 2005a). Therefore, CBA-ID results in data that are sufficiently reliable and have convincing evidence for validity.

CBA-ID COMPARED WITH OTHER ASSESSMENTS

The term *instructional level* is used frequently in education. In fact, there are multiple ways to assess a student's instructional level and there is a multi-million-dollar industry dedicated to doing so. However, many of those measures are problematic for reasons listed below.

Informal Reading Inventories

Betts coined the term *instructional level* in 1946, and unknowingly set in motion an entire industry. Many test publishers have created and sell various informal reading inventories (IRIs) including the Fountas and Pinnell Benchmark Assessment System (F&P; Fountas & Pinnell, 2007), the Basic Reading Inventory (BRI; Johns, 2005), Ekwall Shanker Reading Inventory (ESRI; Shanker & Ekwall, 2002), Qualitative Reading Inventory (QRI; Leslie & Caldwell, 2006), Burns and Roe Information Reading Inventory (B&R; Burns & Roe, 2007), and the Developmental Reading Assessment (DRA; Beavers, 2006).

IRIs vary somewhat in format, but almost all of them involve asking a student to orally read from grade passages while evaluating fluency, followed by answering comprehension questions. The student then reads progressively easier or more difficult passages until the teacher finds the highest level at which the student successfully reads the passage (as judged by acceptable fluency and answering a certain number of comprehension questions correctly). The highest passage at which the student reads successfully is identified as the student's instructional level.

IRIs have tremendous intuitive appeal and are commonly used in schools (Paris, 2002; Paris, Paris, & Carpenter, 2002). Despite widespread use, research on the appropriateness of using IRIs has been minimal at best. In fact, cautions against using IRIs have existed for 40 years (Walker, 1974). Below, we discuss psychometric and practical difficulties with using an IRI to assess students' reading skills.

Despite widespread use, research on the appropriateness of using IRIs has been minimal at best.

Psychometric Difficulties

As described above, tests used for educational decisions should result in data that are sufficiently reliable for a given population and that promote valid decisions. CBA-ID has considerable research regarding reliability and validity, but rarely do publishers of IRIs report the reliability and validity estimates, and those that do often use questionable methods.

Spector (2005) reviewed tests manuals for common IRIs and found that most did not even report reliability. Assessment tools without reported reliability are the equivalent of a psychometric Ouija board; the information that they reveal might be accurate, but we have no way of knowing for sure. If an IRI reports that a student's instructional level is 3.0 grade level, the same test repeated the next day could result in 4.0, 2.4 the day after that, and perhaps as high as 4.6 the following day. Without acceptable reliability, we cannot have confidence in the data.

Many test publishers report correlations with IRI data with other reading measures as evidence for validity. However, correlations do not tell the entire story. Think of a set of scores in which Student 1 scored 98, Student 2 scored 95, Student 3 scored 91, Student 4 scored 88, and Student 5 scored 85. Next, to validate those scores, each student is given a second measure of the same skill and resulted in scores of 65, 63, 60, 55, and 50, respectively. Although the scores were different, the rank order stayed exactly the same, which would result in a very high correlation of .98. One could conclude this correlation as strong evidence for convergence, but what if the criterion for passing both tests was 70? Then, 100% of the students passed the first test, but 0% passed the second test.

We (Parker et al., in press) recently completed a study in which we examined the accuracy of decisions made with the F&P with over 800 students in second and third grade. We compared the F&P level score with a score from the Measures of Academic Progress for Reading (MAP-R; Northwest Evaluation Association, 2004), which is a nationally normed and well-constructed reading measure. The correlation between the F&P reading level and the MAP-R score was .76 for the second graders and .69 for the third graders. Again, these are high correlations, but the data resulted in a consistent decision with the MAP-R only 54% of the time. Therefore, if you wanted to use the data to identify students as needing additional support, you could spend thousands of dollars to purchase the materials, spend hours to train the teachers how to administer the test, dedicate hundreds of hours of instructional time to conducting the assessments, or you could invest 25 cents; simply take a quarter and flip it every time a student enters the door and you will get it right nearly as often.

Practical Difficulties

Besides psychometric considerations, what makes a good educational measure? Glover and Albers (2007) recommended that measures must be cost-effective, aligned with curricula, and time efficient. It would go beyond the scope of this book to review the costs for IRIs, but it suffices to say that IRIs require at least a moderate monetary investment for each assessment kit. Practical considerations are also closely related to time and use of the data. Most IRIs require approximately 20 to even 30 minutes to administer and are completed one-on-one. Thus, a classroom of 30 students would require anywhere from 600 (10 hours) to 900 (15 hours), which equates to 2 or 3 complete days to complete the assessment. Moreover, most schools use IRIs multiple times throughout the year, which could equal as much as 2 weeks of instructional time dedicated to completing the assessment. If the data were instructionally useful, then the time may be well spent, but the data would have to be critically important to warrant that much instructional time, not to mention human resource.

Measures must also be aligned with the curriculum being used. The overlap between major curricula and most IRIs is completely unknown. Moreover, the generalizability of the data is suspect. If a student scores an instructional level of 3.0, for example, the teacher still does not know what 3.0 means. How does that level correspond to other reading material? The F&P is part of a comprehensive system that includes books leveled to F&P levels. However, the publishers include no information about how those books were leveled or how well the score corresponds to the level of the books. We can tell you from experience that just because a student is measured to read at an M level and a book is supposedly written at an M, that does not mean that the student will be able to successfully read the book.

In addition to questions about correspondence between level score and reading level, we question the concept of reading level for individual students. As professionals who have worked with hundreds of students across the country, we frequently conduct various reading assessments with students. One time in particular the first author was leading a data collection effort that involved assessing oral reading fluency with several hundred elementary-age students. Thus, I (M. K. B.) spent an entire day listening to students read from graded passages taken from a highly respected measure of oral reading fluency. The task involved having each student read three separate probes and then recording the median score. Anyone who has conducted a large number of oral reading fluency assessments in 1 day can tell you, it can be a very boring task. I attempted to pass the time by recording the reading rate of students into a Microsoft Excel spreadsheet to see whether there were any differences in words read correctly per minute between boys and girls. Of the three passages, one dealt with dinosaurs and one dealt with cooking (I do not recall the topic of the third passage). At the end of the day, I was quite surprised to learn that the boys read the dinosaurs passage with statistically significantly higher fluency (as measured by words read correctly per minute) than the girls did, and the girls demonstrated significantly better fluency for the cooking passage than did the boys.

Although the finding described above might perpetuate gender stereotypes, it also made an important point. The three passages were supposedly written at the same grade level, but significant differences in readability occurred by gender group within the grade level. How well a child reads a particular passage has more to do with each child's individual background knowledge, vocabulary, and interest than it does the supposed difficulty level with which the book was written, and that information cannot be captured in any individual-level score, be it reported as a grade level (e.g., 3.0) or a readability level (e.g., level M). Data never generalize to or from the individual. If you administered an IRI to 100 students and matched the reading level to their score, you would probably have a decent match 50–66% of the time, but you would have no way of knowing for whom the match would not be aligned. Moreover, the more extreme the reading score, the less likely to match well. Thus, reading levels from IRIs would likely not provide useful information for students who are struggling readers and those who are highly skilled readers, and those are *exactly* the groups for whom we most need accurate data.

IRIs have been around for a long time and are frequently used. However, they have either unknown or generally poor psychometric characteristics, require an excessive amount of time to use, and may not provide useful data, especially for students with extremely low

or high reading skills. Providing reading material that represents an instructional level for an individual student is a critical component of effective reading instruction. Unfortunately, IRIs cannot be effectively used to do so.

Curriculum-Based Measurement

In 1977, Deno and Mirkin proposed CBM as an alternative to IRIs in response to the factors discussed above. CBM administration involves having a student read orally from a given passage for 1 minute and recording the number of words read correctly per minute (WRCM) and the number of errors per minute (EPM). Deno and Mirkin (1977) recommended that students were reading at an instructional level if they read 30–49 WRCM for students in first through third grade, 50–99 WRCM for students in fourth grade and above, and 3–7 EPM for any grade.

There is a wealth of research regarding CBM for reading (CBM-R). Meta-analytic research found high reliability coefficients (Wayman, Wallace, Wiley, Ticha, & Espin, 2007), high correlations with norm-referenced tests (r = .60–.70; Reschly, Busch, Betts, Deno, & Long, 2009), and high correlations with performance on statewide achievement tests (r = .69; Yeo, 2009). Moreover, CBM-R data resulted in over 80% correct classification between the CBM-R data and the MAP-R measure of reading comprehension (Parker et al., in press). Thus, CBM-R seems to be an effective approach to screen students for reading difficulties. It also seems to be ideally suited to monitor reading progress from core instruction or reading interventions (Shapiro, 2011), especially given that the entire assessment would require less than 5 minutes per student. In fact, if the assessor used three passages and recorded the median score, then the assessment would require only 3 minutes and a class of 30 students could be assessed in 1.5 hours, as opposed to the 10–15 hours to conduct an IRI.

Although CBM-R data can be an important component in an assessment-to-intervention model, their utility in determining an instructional level is mostly unknown because very little research has examined this use of CBM-R data. The WRCM and EPM instructional-level criteria suggest that a student in grades 1 through 3 could read 23 of 30 (77%) words correctly to 42 of 49 (86%) words correctly, both of which do not seem to be especially indicative of proficient, or even instructional-level, reading. Moreover, the instructional-level criteria for CBM-R data were established at a school that was part of the precision teaching program being conducted in Minnesota and were not derived from research (S. L. Deno, personal communication, April 15, 2005). Recent research regarding the use of CBM-R to identify an instructional level found that the criteria recommended by Deno and Mirkin (1977) substantially overestimated the reading skills of students by identifying an instructional level when they in fact demonstrated considerable difficulty reading the passage (Parker, Burns, & McComas, 2013).

CBM has many similarities to CBA-ID, but has some fundamental differences. First, CBM is a GOM and CBA-ID is an SMM. Therefore, CBM assesses overall reading skills, but CBA-ID assesses how well a student reads a particular set of materials. CBM attempts to generalize to the broad construct of reading, but CBA-ID makes no assumptions about generalization. Data do not generalize from an individual student (i.e., one student's CBA-

ID data will not generalize to other students, even if similar in many characteristics), but they also do not generalize to the individual student (i.e., CBA-ID criteria are not assumed to generalize to any single student). Much like IRIs, CBM attempts to generalize the score to all written material. We question the generalizability of any one score to the entire universe of reading materials. When we complete a CBA-ID, we have little information about the student's overall reading skills, but we do know how well he or she will interact with the text that will be used for instruction, and we suggest that those data are what most classroom general and special education teachers really want to know.

CBM-R is a well-researched tool that is very useful for screening children and to monitor progress on a more frequent (e.g., weekly) basis because it is highly reliable, corresponds with reading comprehension, is not expensive, and does not require much time to complete. However, much like IRIs, CBM-R data do not seem to provide useful information when determining an instructional level.

Conclusion

Shapiro (2011) presents an assessment-to-intervention model that involves four steps: (1) assessing the academic environment, (2) assessing instructional placement, (3) modifying instruction, and (4) monitoring progress. Assessment data drive all four steps and different assessments serve different purposes. Thus, different assessment approaches are more relevant to certain decisions within an assessment-to-intervention/instruction framework than others. For example, CBM was noted to be the most effective assessment approach for progress monitoring, which seems to be a point of consensus in the literature (Burns, Dean, & Klar, 2004; Deno, 2003; Gresham, 2002; Shinn, Rosenfield, & Knutson, 1989), but CBA-ID may be more relevant for determining how to best modify instruction (Burns, Dean, & Klar, 2004; Shapiro, 2011). The role of IRIs in instruction seems somewhat unclear, but using CBA-ID to determine whether a particular set of materials matches student skill seems superior to making the same judgment with IRI data. Moreover, including CBA-ID along with CBM in Shapiro's integrated assessment-to-intervention model resulted in improved student outcomes (Burns, 2002; Shapiro & Ager, 1992).

ASSESSMENT PROCEDURES

Any assessment procedure should follow a standard administration because if the assessment is not administered the same way every time, then the data cannot be compared across assessments or to a criterion (Kaplan & Saccuzzo, 2001). Thus, clear directions for administration should be outlined so that the administration can be duplicated across times and potential assessment conditions (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 1999). The administration procedures for CBM are well established (e.g., Hosp et al., 2006; Shinn, 1989), but some standardized measures, such as IRIs, include a high level of subjectivity because the assessor rates student behavior. As discussed in Chapter 1, instructional assess-

ment (Gravois & Gickling, 2008) also has some level of subjectivity to it, but using CBA-ID within the model makes the decisions more objectively based.

Assessing the Instructional Level

There are specific administration procedures for all approaches to CBA-ID, and we discuss them in detail for reading, writing, and math in later chapters. All approaches involve using the instructional material as the assessment, directly sampling student performance, timing students as they perform the task, and comparing the data to an instructional-level criterion. Teachers have often questioned the validity of data obtained from timed assessments because they wonder about the negative consequences of having a student watch the clock as he or she works. We have several responses to that concern. First, we time the behavior for 1 minute in reading because that provides a sufficient sample of the behavior. Reading for less time (e.g., 30 seconds) did not result in reliable data but going beyond 1 minute did not improve reliability (Burns et al., 2000). Data for math and writing are converted to a per-minute metric, but students could be timed for any length and data. We suggest providing enough time for the student to complete the task (e.g., 2 minutes for single-digit multiplication and 3 minutes for writing tasks), but to limit it to the smallest interval needed to obtain a valid score. There is little benefit to providing too much time and no need to cut back time so as to pressure the student. Timing the assessment also enhances the standardization of the assessment. Finally, we just do not believe that timing is a big issue for children. We have seen countless examples that demonstrate when educators consider the timing to be routine, students follow the routine as well and become comfortable with it. There may be instances when timing is problematic for a given student (e.g., a student who stutters), and accommodations can be made on a student-by-student basis, but timing should not be an issue for a vast majority of the students.

Assessing the AR

There is one aspect of CBA-ID that remains somewhat constant across academic domains. We assess the AR for math, reading, and spelling the same way. Students can become frustrated if the task is too difficult (e.g., the student can read less than 93% of the words), but student frustration can also be the result of attempting to cover too much information. The appropriate amount of information that a student can manage and maintain while learning a particular skill/lesson is called the AR (Gravois & Gickling, 2002). Previous research found that students with documented behavioral difficulties remained engaged in task-relevant behavior during a word recognition lesson that contained 90% known words, until the lesson exceeded the students' AR, at which time the frequency of off-task behavior more than tripled (Burns & Dean, 2005a).

AR is based on the theory of retroactive cognitive interference, which occurs when students learn a new item, but then cannot recall the new item after learning a subsequent item. In other words, if you want to teach children eight items, but they can only learn four at one time, then they will learn items 1, 2, 3, and 4 with little problem, and not only will

they not learn items 5, 6, 7, and 8 but attempting to teach them will cause students to forget the items that you just taught them. Have you ever seen a student who "knew it one day, but didn't know it the next day"? We believe that one reason why we see memory difficulties and inconsistent performance is that we do not frequently enough consider the limits of human memory, and assessing an AR is one way to do so.

There are potentially multiple ways to assess the AR, but we describe one method in Table 3.1 to ensure standardization. Previous research regarding the reliability and validity of AR data used the same approach (Burns, 2001; Burns & Dean, 2005a; Burns & Mosak, 2005). We begin by identifying a series of known and unknown items, which are discussed in more detail later. We need to identify five known items for children in kindergarten and younger, and eight known items for older children. We then write the known and unknown items on index cards and ask the student to correctly respond to the item (e.g., read the word, tell me the sound that the letter makes, what is 3×3). Those to which a correct response is given are counted as known items and those not correctly responded to or those to which a correct response is given after 2 seconds are counted as unknowns.

After identifying known and unknown items, the unknowns are taught using IR (Tucker, 1989), which is described in more detail in Chapter 7. First, each unknown item is presented to the student while verbally stating the correct response (e.g., "This is 4 times 4, and 4 times 4 equals 16"). Second, the student is asked to restate the correct response (e.g., "4 times 4 equals 16," correctly reading the word, or stating the sound that the letter makes). Third, the unknown item is rehearsed with IR in the following manner: first unknown, first known; first unknown, first known, second known; first unknown, first known, second known, third known; first known, second known, third known, fourth known; first known, second known, third known, fourth known, fifth known. The rehearsal would stop here for kindergarten and preschool children, but would continue in the same manner for older children until all eight known items are presented, which is considered one set.

TABLE 3.1. Procedures for Assessing an Acquisition Rate

1. Write unknown and known items on index cards.
2. Teach the words with incremental rehearsal.
3. Count any error made by the student. Errors are any incorrect response or correct responses after 2 seconds of presentation.
4. Keep adding in unknown items until the student makes three errors while rehearsing any one item. The errors may occur for a target item, a previously taught item, or a known item.
5. After the student makes three errors while rehearsing any one item (a set), stop the assessment.
6. Shuffle the items that were taught and show each one final time. Ask the student to correctly respond to each when you show the card (e.g., read the word, state the answer to a math fact, tell me what sound the letter makes). Items that are correctly responded to within 2 seconds of presentation are considered known, and those that are incorrectly responded to or correctly responded to after 2 seconds of presentation are considered unknown.
7. Count the number of known items, which equals the acquisition rate.
8. Test for retention at least 1 day later by repeating Step 6.

After completing the rehearsal pattern for the first unknown word described above (the first set), a new unknown item would be introduced (the second set), the previous unknown item is then treated as the first known item, and previous final (fifth or eighth) known item is removed from the deck. Any time a student does not correctly respond to an item written on the card, regardless if it is designated as "unknown" or "known," it is immediately corrected and counted as an error. New unknown items are added into the sequence until the student makes three errors while practicing a new item (one set). At this time, the number of unknown items successfully completed is recorded as the AR. For example, if a student rehearses the first four unknown words while making few errors, but makes three errors while completing the fifth word, his or her AR would be 4. As stated earlier, previous research found that ARs can be reliably measured ($r > .90$) for third- and fifth-grade students (Burns, 2001) and were highly correlated ($r = .70$) with a standardized norm-referenced measure of memory (Burns & Mosack, 2005). The specific steps in the assessment procedure are listed in Table 3.1 and are included in Appendix A7.

The role of ARs in intervention and instructional design are discussed in Chapter 7, but generally speaking, the AR provides an estimate about the appropriate number of new items to include when planning instruction. If a student's AR for math facts is 3, then teachers and interventionists know that providing three new math facts in any one lesson or intervention session would likely be sufficient. Teachers know that it is better to include fewer items than too many, but measuring the AR allows for precision in planning so that instructional time and student capacities can both be used to their maximum potential.

CONCLUSION

CBA-ID is an assessment approach with a specific goal. The data obtained from CBA-ID measure of reading, writing, and math can be used to design intervention and to modify instruction, and doing so results in reliable data and valid decisions. CBA-ID has limited utility for other aspects of the instructional process. For example, CBM is a stronger tool to use when monitoring student progress. The information described above demonstrates the utility of CBA-ID as an assessment tool, but it must be perceived as an assessment. Within an assessment paradigm, CBA-ID is less standardized and more informal than many measures commonly used in schools, but informal measures that directly assess the skill being taught result in data that are more useful for formative evaluation decisions. If you want to decide whether a student should be identified as having a disability, if you want to determine whether a student has passed state standards, or if you want to hold a teacher accountable for a test score, then we recommend that you use a different assessment tool than CBA-ID. However, if you want to use data to design instruction, to determine where to target your intervention efforts, or how to modify an intervention or instruction, then CBA-ID is an ideal tool to use. We next discuss how to do so in the chapters that follow.