SECTION III

READING RECOVERY USES STANDARD ASSESSMENT MEASURES.

This section first describes *An Observation Survey of Early Literacy Achievement* (Clay, 1993a/2002) and its history and use in early literacy settings as well as in Reading Recovery. Second, Reading Recovery students' performance in follow-up studies using norm-referenced tests is reported.

A. Reading Recovery uses *An Observation Survey of Early Literacy Achievement,* a standard measure developed in research studies, with qualities of sound assessment instruments with reliabilities and validities and discrimination indices.

1. The Observation Survey was developed to meet the unique need to assess emergent literacy in young children.

In the 1960s there were few studies of literacy acquisition or credible theories of literacy development using close observation over time during the first year of school. Marie Clay was committed to rationality and scientific methodology; the research designs she used and the standard measures she developed followed rigorous standards of research. Since the 1960s, Clay has engaged in more than 40 years of in-depth research and analysis of evidence to construct her theory and validate measures that describe the range of differences and change over time among the lowest-achieving students (Clay, 2001). These studies, discussed in *Observing Young Readers* (Clay, 1982), extend or verify elements of her theory.

Clay sought to understand why children fail to realize their learning potential and to describe the course of literacy development and the different paths students might take. The research methods she used produced assessment tools that have high construct and face validity and high reliability measured in large-scale studies. Measurement error within these tasks is greatly reduced with individual administration and with standardized administration procedures (see *An Observation Survey of Early Literacy Achievement,* Clay, 1993a/2002).[3]

[3] Until the Observation Survey was published in 1993, it was known as the Diagnostic Survey (Clay, 1985).

Tasks included in the Observation Survey incorporate both open and closed tasks. They allow for observation of emerging, tentative behaviors to detect the variability of individual paths to literacy achievement. "Powerful statistical analyses have shown that these procedures, which permit more detailed recording of individual responses than a normative test, nevertheless have proved to be sound measurement devices" (Clay, 1982, p. 6).

A particular strength for Reading Recovery in the United States is that parallel instruments have been developed for the Spanish language, and they have been subjected to the same rigorous analyses for reliability and validity (see Escamilla, Andrade, Basurto, & Ruiz, 1996).

2. The Observation Survey comprises systematic and controlled observation measures to assess young children's literacy knowledge and to detect evidence of progress in the early stages of literacy learning.

The Observation Survey adheres to characteristics of good measurement instruments, namely, a standard task, a standard way of administering the task, ways of knowing about reliability of observations, and a real-world task that establishes validity of the observation.

The six tasks of the Observation Survey have been widely used with the five- to seven-year-old age group; they are not in-house instruments used only for Reading Recovery assessments. Users of the Observation Survey include classroom teachers, teachers working individually with children having temporary difficulties with literacy learning, administrators who want accounts of individual progress of children across time, and researchers probing how young children learn about literacy.

These controlled observation tasks have been widely used in literacy research. They

> can feed data into analyses of researchers, and best of all, they can provide evidence of learning on repeated measurements of tasks the child is actually undertaking in the classroom. *In every way the information that is gathered in systematic observation reduces our uncertainties and improves our instruction.* (Clay, 1993a/2002, p. 2)

The Observation Survey adheres to characteristics of good measurement instruments, namely, a standard task, a standard way of administering the task, ways of knowing about reliability of observations, and a real-world task that establishes validity of the observation.

The Observation Survey is comprised of six literacy tasks with established validity and reliability (see Clay, 1993a/2002).

- Letter Identification (to identify known letters and preferred mode of identification)

- Word Test (to determine whether the child is building reading vocabulary)
- Concepts About Print (to find out what the child has learned about the way spoken language is put into print)
- Writing Vocabulary (to find out whether the child is building a writing vocabulary)
- Hearing and Recording Sounds in Words (to assess phonemic awareness by responses to sound-letter associations)
- Text Reading (to determine appropriate level of text difficulty and to record, using a running record, what the child does when reading continuous text)

The six tasks can be justified by theories of measurement, and they take other theories into account (from the psychology of learning, developmental psychology, studies of individual differences, and theories about social factors and the influences of contexts on learning). Stanines and reliability/validity data are provided for five of the tasks within the Observation Survey.

In the 2002 edition of the Observation Survey, new norming data are provided. New norms will be established for the United States during the 2002–2003 academic year.

Many of the tasks in the Observation Survey are similar to tasks in other widely used standardized and norm-referenced tests. Informal reading inventories are used to determine the appropriate text level for a student's instruction, to monitor progress of individual students, and to obtain pre- and post-test scores in the primary grades. The tasks of the informal reading inventories (word lists and text reading) are similar to the Word Test and Running Records of Text Reading in the Observation Survey.

Word identification tasks are common in both standardized and norm-referenced tests (Woodcock-Johnson III, Slosson Oral Reading Test, Qualitative Reading Inventory-3, Iowa Test of Basic Skills, and others). Reading passages organized along a gradient of difficulty are also prevalent in many tests (Basic Reading Inventory, Qualitative Reading Inventory-3, Gates-MacGinite Reading Tests, and Iowa Test of Basic Skills among others).

Many tests assess beginning readers' phonemic awareness. The Hearing and Recording Sounds in Words task of the Observation Survey requires the student to articulate words slowly, supplying the letter or letters associated with the phonemes within each word of the dictated sentences.

3. Reading Recovery analyses of text reading levels provide descriptive data of behavior on a scale of relative difficulty and provide data about change across time.

The text reading measure used within the Reading Recovery program uses texts from the Scott Foresman Reading Systems Special Practice Books (1979). To standardize the administration of the text reading measure, brief story introductions were produced and a standard format of administration was established. In a 1987–1988 study, Ohio State University researchers piloted texts and procedures. Changes were made and a larger-scale sampling and comparison was completed to test these materials against a previously used testing program, with resulting .85 reliability.

In 1990–1991, a random sample of 155 urban kindergarten and first-grade children were sampled using the Diagnostic Survey tasks[4] (Clay, 1985) and the Reading Recovery text reading materials. An analysis of the text reading materials was completed on the first graders in the study (n=96) to determine the reliability of the scale using a Rasch rating scale analysis (Wright, Linacre, & Schulz, 1989). This analysis showed that the text reading scale had a reliability of .83 (person) and .98 (item). When the text reading measure was combined with two other tasks (a measure of print concepts and phonemic representation), the power of the assessment increased. It verified that while the text reading measure does not provide an equal interval scale, the item difficulty scale itself across these three measures is robust and highly reliable (scale formula r=.99).

Reading Recovery analyses of text reading levels provide descriptive data of behavior on a scale of relative difficulty, and they provide data about change across time. These analyses are appropriate assessment and measurement techniques commonly used in the educational measurement field to validate such observation data as ordinal information, time series samples, and more. These techniques to transform observations into measures date back to the turn of the century (Thorndike, 1904; Thurstone, 1925; Wright & Linacre, 1989) and have since been perfected by advanced statistical procedures (Rasch, 1960, 1980).

4. Reading Recovery appropriately uses the Observation Survey in pre-treatment and post-treatment analyses of children's progress.

With the exception of the Letter Identification task and the

[4] The Diagnostic Survey (1985) was published as part of *The Early Detection of Reading Difficulties: A Diagnostic Survey With Reading Recovery Procedures.* It was later renamed The Observation Survey of Early Literacy Achievement and published separately in 1993 and 2002.

Writing Vocabulary tasks, all sub-tests of the Observation Survey have three or more alternate forms. Data collected annually using the standard assessment measures offer a variety of ways to verify the quality of decisions made about selection of students and outcomes achieved. The standard measures of the Observation Survey are used to select the lowest literacy achievers for service and to make reliable decisions about student progress.

Reading Recovery children begin first grade with much lower scores than their peers; yet children who meet discontinuing criteria reach approximate parity by the end of first grade. Discontinuing decisions are also corroborated by external factors such as classroom teacher perceptions and low rates of retention and placement in special education. The methodology must fit the research questions and issues being addressed. The goal remains the same: select the lowest-achieving students, provide early intervention that reduces the need for further remediation, and monitor change over time using standard measures appropriate for the beginning reader.

B. Reading Recovery students perform well on norm-referenced tests.

The Internet letter suggests use of norm-referenced tests that are widely available and commonly used in reading intervention research. Although traditional standardized tests may yield valid comparisons of mean scores derived from groups of students who are already reading, they are not sensitive to variability in emerging knowledge. These tests are not useful as baseline measures for assessing change over time in individual young learners. For this reason, such tests are not used for selection into Reading Recovery.

Some studies, however, have examined Reading Recovery children in Grade 1 (see for example Pinnell et al., 1994) using such assessments as the Slosson Oral Reading Test and the Woodcock-Johnson III. Many more studies have used standardized independent measures and state assessments to explore Reading Recovery children's performance in subsequent grades.

In California, a study of four cohorts of former Reading Recovery students (760 students in Grades 2, 3, 4, and 5) revealed that 68–85% (percentages varied by year and by group) scored at Stanine 4 or above on both of two high-profile standardized tests, the Iowa Test of Basic Skills and Stanford Achievement Test 9 (Brown et al., 1999).

Two longitudinal studies in Texas used the Gates-MacGinitie Reading Test and the Texas Assessment of Academic Skills to

explore subsequent performance of Reading Recovery children compared with a random sample of their classroom peers (Askew et al., 2002). Findings showed that 80–85% of the former discontinued Reading Recovery children passed the fourth-grade Texas Assessment of Academic Skills reading test and 90–93% passed the writing test. Annual progress on the Gates-MacGinitie Reading Test closely paralleled the progress of the random sample.

An Indiana study of the performance of former successful Reading Recovery children on the Gates-MacGinitie Reading Test found that 86% of those children currently in second grade scored within the average range of scores as established by a random sample group; 84% currently in third grade and 80% in fourth grade scored within the average range (Schmitt & Gregory, 2001). Scores on the third-grade Comprehensive Test of Basic Skills-5/ Terra Nova Form B assessments administered statewide approximated a normal distribution for former Reading Recovery children with a mean at the 45th percentile.

Although the Internet letter suggests use of norm-referenced tests instead of the Observation Survey, these traditional standardized tests are not useful as baseline measures for assessing change over time in individual young learners.

C. Summary

Reading Recovery uses the measurements published in *An Observation Survey of Early Literacy Achievement* (Clay, 1993a/2002), a standard measure developed in research studies with qualities of sound assessment instruments with reliabilities and validities and discrimination indices. The survey adheres to characteristics of good measurement instruments: namely, it is a standard task, there is a standard way of administering the task, there are ways of knowing about reliability of observations, and it is a real-world task that establishes validity of the observation.

Reading Recovery appropriately uses the Observation Survey in pre-treatment and post-treatment of children's progress. Although the Internet letter suggests use of norm-referenced tests instead of the Observation Survey, these traditional standardized tests are not useful as baseline measures for assessing change over time in *young* learners. For this reason, these tests are not used for selection of children for Reading Recovery. When children have developed more literacy skills by the end of Grade 1 and in later grades, standardized measures are often used to examine subsequent literacy achievement for Reading Recovery students (Askew et al., 2002; Brown et al., 1999; Pinnell et al., 1994; Schmitt & Gregory, 2001).