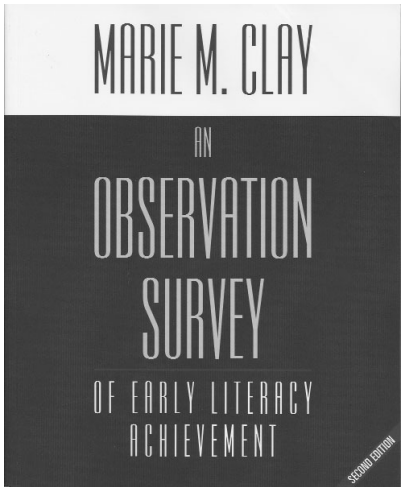

An Evaluation of *An Observation Survey of Early Literacy Achievement*

Robert L. Johnson, Associate Professor, Educational Psychology, University of South Carolina

Jennifer E. Young-Hubbard, Teacher Leader, South Carolina Department of Education

The publication *An Observation Survey of Early Literacy Achievement* (Clay, 2002) is central to the measure of student progress in Reading Recovery and is widely used by teachers of young children to observe and measure the progress of students in the classroom. The Observation Survey was designed to provide a way to observe progress in learning to read during a period of a child's literacy development in which progress is often difficult to document. According to Clay (2002), "It is difficult and perhaps impossible to design good measurement tools for use close to the onset of instruction. Standardised tests…do not discriminate well until considerable progress has been made by many of the children" (p. 11). Early literacy educators need an instrument that provides information about what young children are learning about and learning to do in beginning reading and writing.

This article reviews the Observation Survey to examine its effectiveness in supporting teachers in systematically assessing early literacy achievement and making decisions about student progress. The article brings together the distinct perspectives of two educators. Jennifer is a teacher leader and provides a view from within the Reading Recovery community. Robert is an associate professor in research and measurement and brings the perspective of an assessment practitioner. The authors frame their evaluation of the Observation Survey using three characteristics of assessments that typically are the focus of critical reviews. Those characteristics are the norms, the reliability of scores, and the validity of decisions based on the instrument results. Before discussing these criteria, we examine the overall purpose of the Observation Survey.

Purpose of the Observation Survey

The Observation Survey provides a set of standardized tasks that allow the teacher to observe the young child as he or she engages in reading and writing. "What I like about observation is that it allows us to watch the child as he works, to see at least part of the focus of his attention, to watch him search for information in print and for confirmation of what he thinks" (Clay, 2002, p. 3). The scores on each of the tasks provide quantitative data in order to make comparisons between children as well as to measure the progress of individual children over time. In addition, observation of the child at work on the tasks provides qualitative information concerning the child's strategic operations on print. For example, a child's attempts to monitor reading, to search for information when encountering a difficulty, and to self-correct errors that interfere with the reading process provide the teacher with clues about the child's thinking. Teachers use this information to plan an intervention program such as Reading Recovery or to guide instructional decisions for small groups of children in the regular classroom.

Norms

The Observation Survey provides several types of scores—raw scores, percentile ranks, and stanines—for the tasks of Letter Identification, Concepts About Print, Clay Word Reading,[1] Writing Vocabulary, and Hearing and Recording Sounds in Words. Percentile ranks and stanines are norm-referenced scores that allow

[1] Clay Word Reading is analogous to the Ohio Word Test. See pages 91–95 of *An Observation Survey of Early Literacy Achievement* (Clay, 2002).

comparisons between a student's performance and a sample of students referred to as a norm group. Percentile ranks range from 1 to 99 and indicate a student's ranking on a task as compared to the students in the norm group. For example, a student in the 30th percentile scored as well or better than 30% of the students in the norm group. Stanines are less refined units that cluster scores into 9 levels where 1 is low, 5 is average, and 9 is high. Stanines 1, 2, and 3 correspond to the 1st to 22nd percentile ranks; 4, 5, and 6 correspond to the 23rd to 76th percentile ranks; and 7, 8, and 9 correspond to 77th to 99th percentile ranks.

This article brings together the distinct perspectives of two educators: Robert L. Johnson is an associate professor in research and measurement, and Jennifer E. Young-Hubbard is a Reading Recovery teacher leader.

The norm group is often, but not always, a national sample of students. Norming data provided for the tasks in *An Observation Survey of Early Literacy Achievement* reflect national norms for students in New Zealand schools. National norms allow a teacher to compare the score that his students received on a test to the scores earned by students across the nation (i.e., students in the norm group). For example, a 5-year-old student's score at a stanine of 6 on Concepts About Print indicates to a teacher that the student performed better on this task than a majority of 5-year-old students in the New Zealand norm group. The Observation Survey also contains raw scores and stanines for first-grade students in the United States for the following tasks: Letter Identification, Concepts about Print, Ohio Word Test, Writing Vocabulary, Hearing and Recording Sounds in Words, and Text Reading.

The characteristics of the students in the norm group are critical in allowing teachers to make appropriate interpretations of students' performances. Characteristics that are important include the age level of students, their gender, their ethnicity, and their socioeconomic level. With this type of information, a teacher can decide whether the norm group for a particular instrument provides a relevant comparison for her students. To continue the previous example, students who were age 5 when the test was normed provide an appropriate group for making relevant comparisons with the 5-year-old student tested with Concepts About Print. A norm group composed of second-grade students (age 7) would not be relevant for making comparisons with students who are age 5.

The students in the norm group typically are a representative sample of a larger population of students. For example, the first-grade students who comprise the norm group for a national achievement test are only a sample of all possible first-grade students. The representativeness of the norm group in terms of the larger student population is important in the interpretations of the percentile ranks and stanines. For example, in a norm group the percentage of students by gender should be similar to those in the larger population. If the assessment provides technical information that compares the characteristics of the norm group with the larger student population, then the teacher can determine if the norm group is representative.

The second edition of *An Observation Survey of Early Literacy Achievement* contains national norms based on 796 children between the ages of 5 and 7 in New Zealand. Four students per school were randomly selected from a representative sample of 199 schools selected by the Ministry of Education. The sample represents four age levels: 5.00–5.50, 5.51–6.00, 6.01–6.50, and 6.51–7.00. The decision to report stanines for 6-month age spans is research-based: evidence from the standardization demonstrated that changes in literacy skills occur at a rapid pace for students ages 5–7. Thus, instead of providing stanines for one-year spans, *An Observation Survey of Early Literacy Achievement*

provides stanines for 6-month intervals to allow more frequent observations of student progress. The norm group is described as 52% female and 48% male; however, no other characteristics are reported. In future editions of *An Observation Survey of Early Literacy Achievement,* the addition of tables that report other characteristics for the student population in New Zealand schools and the characteristics of the norm group would provide additional information about the representativeness of the norm group.

Norms for first-grade students in Ohio are included in Appendix 3 of the second edition of *An Observation Survey of Early Literacy Achievement* because comparisons of students in the United States with students in the New Zealand norms are of limited use to teachers and administrators in the United States. The Ohio norms are based on samples of urban children in Ohio during the period of 1990–1991. Norms are reported for three time periods: autumn, mid-year, and spring. The size of the norm groups vary by time of the year and task; group sizes range from 104 to 155 for the autumn, 73 to 114 at mid-year, and 107 to 114 in the spring. Updated norms for the United States currently are being developed and will be based on a large, national, stratified random sample. The current project to update the U.S. norms should improve the relevancy and representativeness of the 1990–1991 norm group reported in the appendix of *An Observation Survey of Early Literacy Achievement.*

Validity

Teachers assess students in order to make decisions about the types of learning experiences that students need as well as the progress they are making. Clay (2002) provides an example of the role of decision making when she states that systematic observations provide "ways of knowing when we can…make valid comparisons" (p. 12). The study of the validity of an instrument such as the Observation Survey allows teachers and literacy researchers to gauge the accuracy of the decisions that they are making based on the assessment. Formally, validity examines "the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores" (Messick, 1993, p. 13). Validity evidence can take many forms. Validity evidence provided in *An Observation Survey of Early Literacy Achievement* includes (a) the extent that the instrument reflects current conceptualizations of literacy, (b) the similarity of students' scores on the instrument with those on other literacy assessments, (c) the ability of the assessment to predict students' scores at a later point, and (d) the expected changes in scores as a result of age, instruction, and experiences with literacy. Each of these aspects of validity evidence is discussed below.

Current Conceptualizations of Literacy

Every assessment instrument, indeed any instructional approach or teaching act, reflects a theoretical construct—a set of assumptions about whatever is being taught or assessed. A literacy assessment is said to show evidence of construct validity if it reflects current theoretical conceptualizations about literacy. By creating an observational instrument with six different tasks, Clay suggests that reading and writing are reciprocal, multifaceted processes, not explained by the simple accumulation of items such as letters, sounds, and words. The Observation Survey provides teachers with opportunities to observe children in the act of constructing meaning from and with texts and to observe each child's system for operating on text as the system itself is being constructed by the child.

In contrast, the reading assessments available when Clay conducted her early research were often single measure tests, or tests which required that a child already be able to read to answer a series of questions. Such assessments reflected a theory that Clay felt could not explain the complexity of what happens when good readers read. She wrote, "A theory of reading continuous texts cannot arise from a theory of word reading because it involves the integration of many behaviours not studied in a theory of word reading" (1993, p. 7). To rely on a single measure of literacy such as word reading or phonemic awareness requires one to assume that the accumulation of enough words or letters and their sounds will eventually lead to proficient reading. Clay rejected this assumption because she felt that it did not sufficiently explain what happens during the process of reading.

A view of reading and writing as complex, reciprocal processes, requiring the construction of meaning and the integration of information from the text as well as from the reader's own stores of knowledge and experience, is now widely accepted by reading researchers and professionals in the literacy community (e.g., Anderson, Hiebert, Scott, & Wilkerson, 1985; Braunger & Lewis, 1985; International Reading Association & National Council of Teachers of English, 1996; Rhodes & Shanklin,

Table 1.
Intercorrelations of Observation Survey Tasks

Tasks	Range of Correlations
Text Level	.47-.89
Letter Identification	.47-.83
Concepts about Print	.56-.79
Clay Word Reading	.59-.89
Writing Vocabulary	.48-.89
Hearing and Recording Sounds in Words	.65-.89

Summarized from Clay (2002, p. 158)

1993). Thus, we conclude that the Observation Survey reflects current conceptualizations of the complexity of literacy learning.

Similarity of Scores on Literacy Instruments

Additional evidence to support the validity of the decisions that teachers make about students based on a literacy assessment is provided when students' scores are similar on other literacy instruments. For example, students' performances on the Text Reading task in the Observation Survey are likely to be similar to their scores on another literacy assessment, such as an informal reading inventory. The similarity of scores across assessments is generally expressed as a statistic referred to as a correlation. In the study of validity, the correlation is expressed as a validity coefficient that ranges from 0.0 to 1.0. A correlation of 0.0 indicates no similarity on students' scores on the two assessments, whereas a correlation of 1.0 indicates that students perform the same on both assessments. McDaniel (1994) provides a general guideline that instruments with validity coefficients of .40 or higher may provide useful information for making decisions about students. As discussed below, the tasks in the Observation Survey consistently are associated with validity coefficients that exceed McDaniel's criterion.

Table 1 summarizes the validity coefficients (i.e., correlations) between the various tasks of the Observation Survey. Correlations between the tasks range from .47 to .89, providing evidence that students perform similarly across the various measures of literacy. Thus, if a teacher finds that a student faces challenges in reading based on one of the Observation Survey tasks, the correlations indicate that this finding will likely be supported by the student's performance on the other tasks. A correlation below 1.0, however, indicates that each task provides unique information about a student's literacy skills, thus reliance on any one task to make decisions about a student would be inappropriate. As Clay (2002) states, "When important decisions are to be made we should increase the range of observations we make in order to decrease the risk that we will make errors in our interpretations" (p. 12).

Additional validity evidence is provided in the form of correlations between student scores on Observation Survey tasks and other literacy assessments. Table 2 summarizes validity information contained in the table "Time Lapse From Entry to School" (Clay 2002, p. 159) and information provided to the authors by Clay.[2] In the first column of Table 2, "Test of Reading Progress" refers to a test created by Clay (1966) that consists of the 15 words from the Clay Word Reading task and the first 30 words of the Schonell R1 developed by Fred Schonell (cited in Buros, 1975, p. 1647). The Schonell R1, also referred to as the Graded Word Reading Test, is the first of seven levels of a reading test developed from 1942–1955 (Buros, 1975, p. 1647).

Combination of the Schonell R1 and Clay Word Reading provided a continuous spread of scores that was suitable for showing differences in groups of children at varying achievement levels during early literacy learning. Clay used correlations between the Test of Reading Progress and the Observation Survey to collect concurrent validity evidence (i.e., correlations between two tests given at approximately the same time). Correlations between the two tests range from .67 to .93, indicating that

[2] Clay indicated that column headings in the table on page 159 will be corrected when *An Observation Survey of Early Literacy Achievement* is reprinted.

Table 2.
Correlations Between Tasks in the Observation Survey and Other Literacy Instruments

	Test of Reading Progress	Schonell R1		Fieldhouse Reading	
Observation Survey Tasks		Age 7	Age 8	Age 7	Age 8
Letter Identification	.83	.86	.81	.80	.83
Concepts About Print	.79	.73	.64	.69	.70
Clay Word Reading	—	.90	.80	.88	.83
Text Reading Accuracy	.93	.80	.69	.77	.72
Error Rate on Text	.85	.78	.77	—	—
Self-correction	.67	.61	.60	—	—

Summarized from Clay (2002, p. 159)

students' scores are related for these assessments of literacy.

Prediction of Scores on Other Literacy Assessments

The ability to predict students' scores on later assessments also provides validity evidence. Shown in Table 2 are correlations between the Observation Survey tasks and Schonell R1 and the Fieldhouse Reading Test, the latter being a reading test developed by A. E. Fieldhouse in 1952 (cited in Buros, 1975, p. 1685). The correlations show the relationship between students' scores at age 6 and their scores at ages 7 and 8. At 7 years of age, the correlations range from .61 to .90 across the assessments. At age 8, the correlations range from .60 to .83. These validity coefficients indicate that students' performance on the Observation Survey at age 6 is predictive of their literacy scores one and two years later.[3] In future editions of *An Observation Survey of Early Literacy Achievement,* the inclusion of a description of the literacy assessments used in the validity studies would support researchers in drawing conclusions about the expected correlations between Observation Survey tasks and the Schonell R1 and the Fieldhouse Reading Test.

Expected Changes in Scores

Additional validity evidence is found in the reporting of task mean scores across the various age groups presented in Table 3. In this form of validity evidence, Messick (1993) indicates that "criterion groups are identified that are expected to differ with respect to the construct being measured" (p. 55). The age groups in Table 3 reflect the length of students' exposure to school. Thus, as the age group increases, students would be expected to differ on the literacy tasks. More specifically, the additional years of literacy instruction would be expected to result in greater mean scores on each of the assessments. This is the case across tasks, providing additional validity evidence.

In summary, the theoretical rationales and empirical evidence in *An Observation Survey of Early Literacy Achievement* provide support for the appropriateness of teachers' decisions about their students' progress in learning to read. The observation system reflects current conceptualizations of literacy. In addition, *An Observation Survey of Early Literacy Achievement* offers empirical evidence supporting interpretations of students' reading progress. Additional evidence of validity includes the similarity of students' scores from the Observation Survey with those on other literacy assessments, the ability of the assessment to predict students' scores at a later point, and the expected changes in scores as related to age, instruction, and literacy experiences.

[3] It should be noted that these correlations were based on the scores of students randomly selected from the general population for the purpose of norming the Observation Survey. Interventions such as Reading Recovery are intended to alter the trajectory of student learning and therefore change such predictions of student progress.

Jennifer Young-Hubbard works with a Reading Recovery student. The Observation Survey supports teachers and teacher leaders in systematically assessing early literacy achievement and making decisions about progress of emergent readers.

Reliability

Clay (2002) indicates that assessments are needed in the classroom that provide "systematic observations of children who are in the act of responding to instruction, observations that are reliable enough to compare one child with another, or one child on two different occasions" (p. 3). She notes that "an unreliable test score means that if you took other measures, at around the same time or at another time, you might get very different results" (p. 13).

A basic question for the teacher, then, is whether a student is likely to earn the same score on a task if the teacher were to re-administer the assessment without intervening instruction. If student scores are consistent, a teacher can rely on those scores to make decisions about students. If scores vary from one administration to the next, then the teacher would be unsure about what decision might be appropriate for students. As Clay (2002) states,

> We have to be concerned with whether our assessments are reliable because we do not want to alter our teaching, or decide on a child's placement, on the basis of a flawed judgment. We need to be able to rely on the data from which we make our judgments (p. 13).

All assessments are unreliable to some extent. Intuitively, teachers understand that a student's reading performance may vary somewhat from one text to another. A child's score on the Writing Vocabulary task may be slightly higher from one occasion to the next. Although assessments are unreliable to some extent, it follows, then, that these same assessments are reliable to *some degree.* Teachers understand, for example, that students who score high on the Concepts About Print task administered using *Follow Me, Moon* (Clay, 2000) are likely to score similarly if *Sand* (Clay, 1972) or *Stones* (Clay, 1979) were used for administration of the task.

Table 3.
Means Scores for Observation Survey Tasks Across Age Groups

	Age Groups			
Observation Survey Tasks	5.00-5.50	5.51-6.00	6.01-6.50	6.51-7.00
Letter Identification	39.0	46.6	50.7	51.6
Concepts About Print	13.5	15.5	18.0	18.7
Clay Word Reading	4.3	7.7	11.4	13.0
Writing Vocabulary	12.9	23.8	42.7	51.0
Hearing and Recording Sounds in Words	15.6	23.6	30.7	33.2

Summarized from Clay (2002, pp. 149–152)

Table 4.
Reliability Estimates for Observation Survey Tasks

Assessments	Test-Retest	Internal Consistency: Alpha	Internal Consistency: KR-20	Internal Consistency: Split Half
Letter Identification	—	.95	—	.97
Concepts About Print (4 forms)	.73–.89	.78, .87	.85	.84–.95
Clay Word Reading (3 forms)	—	.92	.90	—
Writing Vocabulary	.62, .97	—	—	—
Hearing and Recording Sounds in Words	.64	.96	—	.84–.88

Summarized from Clay (2002, pp. 160–161)

To estimate the reliability of an assessment like the Observation Survey, researchers like Clay conduct reliability studies to examine the consistency of students' scores across various conditions. Reliability studies examine the consistency of students' scores (a) on different items within a single assessment (internal consistency), (b) on different test forms that assess the same content (parallel forms), (c) on the same assessment across different occasions (test-retest), and (d) assigned by different observers (interrater reliability). The level of consistency is reported as a statistic, referred to as a reliability estimate, that ranges from 0.0 (no consistency) to 1.0 (perfect consistency).

Several sources provide guidance in the interpretation and evaluation of reliability estimates. Some authors indicate that research studies and low-stakes assessments require a minimal reliability of .70; whereas, in applied settings with high stakes, tests require a minimal reliability of .90 (Herman, Aschbacher, & Winters, 1992; Nunnally, 1978). As seen in Table 4, studies of the internal consistency of the various assessments in the Observation Survey report reliability estimates that range from .78 to .97. The lowest internal consistency estimates are associated with the Concepts About Print and Hearing and Recording Sounds in Words assessments. Reliability estimates for Concepts About Print range from .78 to .95; however, a majority of estimates are in the .80s. Hearing and Recording Sounds in Words estimates range from .84 to .96. Internal consistency estimates for Letter Identification and Clay Word Reading range from .90 to .97; these are respectable levels.

Test-retest is important when teachers "want reliable ways to compare a student on two of his own performances" (Clay, 2002, p. 12). Reliability estimates for scores on the Observation Survey tasks across two occasions range from .62 to .97. Writing Vocabulary has the lowest estimate (.62) in one reliability study and the highest estimate (.97) in another study. Hearing and Recording Sounds in Words has the next lowest test-retest reliability estimate of .64. Concepts About Print estimates range from .73 to .89.

Reliability also is a factor in determining a student's text reading level. Text reading level is determined by calculating the highest text level read by a child at or above 90% accuracy. This calculation is reliable only to the extent that a teacher's running record of a child's reading would be consistent with another teacher's recording of the same child's reading. This form of reliability, referred to as *interrater reliability,* was examined using running records (Clay, 2002, p. 161). No specific reliability estimates are reported; however, Clay indicated that chi-square tests showed no significant differences for raters' recording and scoring of error and self-correction rates, indicating that observers were consistent in their recording.

No estimates are reported for parallel form reliability. Given that Concepts About Print and Clay Word Reading have multiple forms, a study of the reliability across the four forms of Concepts About Print and three forms of the Clay Word Reading seems appropriate. Such a study would allow teachers to examine whether student scores are likely to be consistent across parallel forms of the tasks.

Conclusion

This review examined the norms, reliability, and validity of the Observation

Survey. In the publication *An Observation Survey of Early Literacy Achievement,* Clay provides six standardized tasks and norm-referenced scores to allow comparisons of a student's progress with a national sample of students. Recent norms for New Zealand students are included in the 2002 edition. A project currently underway by the North American Reading Recovery Trainers Group and the National Data Evaluation Center will soon provide updated national norms on a much larger sample than was previously available for the United States.

Clay has provided extensive evidence to support the validity of the Observation Survey. Reliability studies indicate that scores from the Observation Survey provide reasonably consistent information to support teachers in making appropriate decisions about students. Overall, the assessment demonstrates many of the qualities required of an effective early literacy instrument. The Observation Survey supports teachers in systematically assessing early literacy achievement and making decisions about the progress of emergent readers.

References

Anderson, R., Hiebert, E., Scott, J., & Wilkerson, I. (1985). *Becoming a nation of readers: The report of the Commission on Reading.* Washington, DC: National Institute of Education.

Braunger, J., & Lewis, J. (1985). *Building a knowledge base in reading.* Portland, OR: Northwest Regional Educational Laboratories.

Buros, O. (Ed.). (1975). *Reading: Tests and reviews II.* Highland Park, NJ: Gryphon Press.

Clay, M. M. (1966). *Emergent reading behaviour.* Unpublished doctoral dissertation, University of Auckland, New Zealand.

Clay, M. M. (1972). *Sand – the Concepts About Print test.* Auckland, NZ: Heinemann.

Clay, M. M. (1979). *Stones – the Concepts About Print test.* Auckland, NZ: Heinemann.

Clay, M. M. (1993). *Reading Recovery: A guidebook for teachers in training.* Portsmouth, NH: Heinemann.

Clay, M. M. (2000). *Follow me, moon.* Auckland, NZ: Heinemann.

Clay, M. M. (2002). *An observation survey of early literacy achievement* (2nd ed.). Portsmouth, NH: Heinemann.

Herman, J., Aschbacher, P., & Winters, L. (1992). *A practical guide to alternative assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

International Reading Association & National Council of Teachers of English (1996). *Standards for the English language arts.* Newark, DE: International Reading Association.

McDaniel, E. (1994). *Understanding educational measurement* (2nd ed.). Madison, WI: WCB Brown & Benchmark.

Messick S. (1993). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education.

Nunnally, J. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Rhodes, L., & Shanklin, N. (1993). *Windows into literacy: Assessing learners K–8.* Portsmouth, NH: Heinemann.