New Stanines for the Observation Survey: The U.S. Norming Process, 2002–2003

Francisco Gómez-Bellengé, Director, National Data Evaluation Center

An Observation Survey of Early Literacy Achievement (Clay, 1993, 2002) is a standardized assessment with six tasks designed to measure early reading and writing behaviors. Published standardized assessment measures such as the Observation Survey are often normed to allow for the interpretation of a given score on any measure for comparative purposes. For example, one could say that a child's raw score on the Concepts About Print task in fall of first grade was 20. This would be somewhat meaningful to Reading Recovery teachers because they administer this task to many children and they would have an idea of what that performance means relative to other students. They would know this is a relatively high score but not exactly how high. An outsider would have no way of knowing what the score means in a comparative sense. Norming the Concepts About Print task for fall of first grade solves these two problems. It allows both the teacher and the outsider to know that this score of 20 places that child in the 90th percentile rank; that is, this child did better than 90% of the students assessed in fall of first grade.

What exactly, then, are norms? Norms are important statistical data that allow comparisons of local populations of students to a broader-based reference group. The following definition is used by the College Board:

> A statistical description of the test performance of a well-defined group that serves as a reference by which to gauge the performance of the other individuals who take the test. Most norms tables show, in descending order, various test scores and the percentage of people in the reference group who scored below each score level. Thus, knowing an individual's score, you can quickly determine how he or she compares with the reference group (College Board, 2005).

For the U.S. norming process carried out in 2002–03 for the six subtasks of the Observation Survey, the *well-defined group* is first-grade children in mainstream classrooms of schools served by Reading Recovery during that school year. A *statistical description* usually includes

- the average score for a group of students at a point in time
- the standard deviation and/or the standard error (the variation around that score)
- the range of scores (minimum, maximum)
- percentile ranks
- stanines (sometimes computed to facilitate teachers' interpretation of norms)

For Reading Recovery teachers who use the Observation Survey on a daily basis, it may seem strange to think that the Observation Survey is a standardized assessment in the same way the SAT might be. It is, and it is not! Clay defines systematic observation with good measurement instruments as having the following characteristics:

- a standard task
- a standard way of administering the task (Clay, 2002, p. 12).

This is very similar to how the developers (Hoover et al., 2003) of the Iowa Test of Basic Skills talk about standardized tests:

> There is a common myth held by many educators that standard procedures apply to commercially-developed norm-referenced tests but not to criterion-referenced tests … However, standard procedures are a critical prerequisite for nearly every sound use of assessment results… . (p. 8)

In the first edition of *An Observation Survey of Early Literacy Achievement* (Clay, 1993), Clay reported stanines (one type of statistical interpretation) for 300 urban New Zealand children from 1968 as well as about 100 Ohio urban children from 1990–1991 (pp. 88–89). In the second edition of the Observation Survey (Clay, 2002), Clay reported new stanines and norms for New Zealand on the basis of data collected in 2000 on a national sample of 796 children. However, the U.S. was without new normative data for the Observation Survey.

In 2002–03, the National Data Evaluation Center (NDEC) conduct-

ed a study to develop new norms for the U.S. The new stanines are available on the publications page of the NDEC Web site (www.ndec.us) in a document titled "Procedures for Administering Leveled Text Reading Passages and Stanines for the Observation Survey" (Fried, Briggs, Konstantellou, & Rosser, 2005). These new norms are being used for the first time during the current school year, 2005–2006. The full norms are in a document titled "U.S. Norms for Tasks of An Observation Survey of Early Literacy Achievement" (Gómez-Bellengé & Thompson, 2005). What follows is a discussion of how the normative data obtained for the Observation Survey applies to the normal distribution and the implications the new stanines have for student selection in Reading Recovery.

The Observation Survey and the Normal Distribution

The Observation Survey is a standardized test, but it is not considered to be what is commonly referred to as a Published Norm-Referenced Test (PNT). The key difference is that PNTs are designed to have normal distributions, with large numbers of students scoring in the middle ranges and few in the upper and lower ranges. This is convenient when performing statistical computations and is accomplished by removing or adding test items until the range of scores conforms to a normal distribution.

An example of a normal distribution is found in Figure 1. In the normal distribution, stanines 4–6 correspond to an average range in the distribution of scores. Stanines 4–6 correspond to the 23rd through the 76th percentile ranks, which is roughly the middle 50%. Tests whose scores distribute normally are a convenient way of ranking students, that is, of separating students into below average, average and above-average categories.

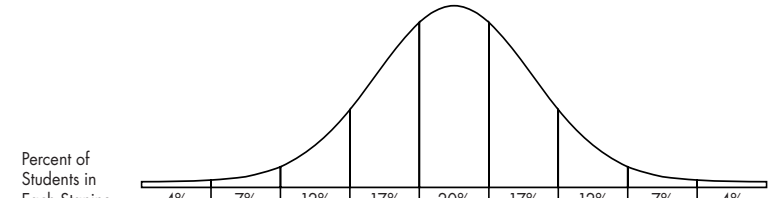
With emerging readers, it is important to know if they fall within an average range or if they are struggling. In providing early literacy intervention, distinguishing between an average reader and a reader that is above average, or between an above average reader (e.g., stanine 7, 77th to 88th percentile ranks) and one that is well above average (stanine 9, above the 95th percentile) is not useful. This is why the tasks of the Observation Survey were not designed to distribute normally.

One could take a task such as Letter Identification and make it distribute normally for any given age. While it is obvious that most adults (or 7-year-olds for that matter) recognize all the letters of the alphabet, one could make a number of adults do poorly on the test by, for example, reducing the length of time given to answer all the items (again, for example) to 50 seconds for the 54 letter forms. One could even get creative and use unusual fonts or handwritten letters. This would be convenient for a statistician trying to create stanines or measure change over time, but it is not very meaningful when you need to know how many 7-year-olds can identify 54 letter forms in a reasonable time frame.

If you allow a 7-year-old or an adult a couple of minutes to identify the alphabet, most will get it 100% right. If you ask 3-year-olds, close to 0% will get it right. How useful is a test where most of the takers score 0% or 100%? This depends on the purpose of the assessment. If the assessment is meant to yield a range of scores for 3- or 7-year-olds, the Letter Identification task is a poor assessment. If the purpose of the assessment is simply to determine whether a child knows all the letter forms of the English alphabet, it works quite well.

In Reading Recovery, using the same tasks to assess children before and after the intervention is essential. This means a task that could distribute normally at one point in time will not at another point in time. This is not a problem because the purpose of assessment in Reading Recovery is not to compare one child to all others—that is to rank their performance relative to a group—but

Figure 1. Example of a Normal Distribution



Percent of Students in Each Stanine	4%	7%	12%	17%	20%	17%	12%	7%	4%
Stanine	1	2	3	4	5	6	7	8	9
Percentile Ranks	below 4	4–10	11–22	23–39	40–59	60–76	77–88	89–95	above 95
	Below Average			*Average*			*Above Average*		

to compare their performance to their own previous performance and to determine whether they have reached a certain level on a given task. In this sense, the post-intervention assessment in Reading Recovery is criterion referenced. Although specific cut scores are not used for making decisions about Reading Recovery children, more than one score is examined and the teachers' observations of actual reading behavior also matter.

Adjusting the Stanines for Student Selection

When the Text Level Reading task is administered to struggling early readers in fall of first grade, most score a 0 or 1, that is, most are at or near the bottom of the range. This is called a *floor effect*. When the Letter Identification task is administered to first graders in spring, most identify 52, 53, or 54 letter forms; most are at or near the top of the range of possible scores. This is called a *ceiling effect*. This happens when the tasks are a closed set, with a pre-set maximum score.

Put another way, "… some skills, such as alphabet knowledge, concepts of print, and phonemic awareness, are universally mastered in relatively brief developmental periods. As a consequence, the distributions of data from these variables are skewed by floor and ceiling effects that, in turn, influence the correlations used to establish reliability and validity of assessments" (Paris & Hoffman, 2004).

The fact that various tasks of the Observation Survey exhibit ceiling or floor effects reflects the nature and purpose of this assessment; it is a characteristic of the Observation Survey, not a flaw. When stanines are computed for the Observation Survey, we end up with empty stanine categories due to the floor and ceiling effects. This is why the Reading Recovery North American Trainers Group has decided to use modified stanines for the student selection process. They were first published in "Procedures for Administering Leveled Text Reading Passages and Stanines for the Observation Survey" (Fried et al., 2005) and are reproduced in table form on the following pages.

For the Letter Identification task, as stated earlier, most first-grade children know most letter forms by year-end. Many children also know most letter forms in fall and at mid-year. This is why there are no raw scores corresponding to stanines 7, 8, and 9. Another way of saying this is that nothing is exceptional about a first-grade child who recognizes all letter forms in fall or in spring. However, a child who recognizes fewer than 45 letters in fall (the low end of stanine 4) is below the average range and a child who recognizes fewer than 39 letters (the low end of stanine 2) in fall is well below the average range. This is useful information for teachers, even though the distribution of scores is not normal.

For the Ohio Word Test, Writing Vocabulary and Text Reading Level tasks, floor effects led to the adjustment of the stanines for the purpose of student selection. The unadjusted stanine for Ohio Word Test scores 0, 1, and 2 is stanine 3, with no raw scores corresponding to stanines 1 and 2. It is simply more convenient, when selecting students for Reading Recovery, to distinguish between Ohio Word Test raw scores of 0, 1, and 2. This was done by assigning the raw score 0 to stanine 1, raw score 1 to stanine 2, and raw score 2 to stanine 3.

Comparing the New and Old Norms

Student samples upon which the U.S. norms for the Observation Survey were derived in 1990–1991 and again in 2002–2003 are not strictly comparable. The old norms were derived from a sample of urban children (from the midwest). The new norms are taken from a national proportional stratified sample that approximates the entire population of U.S. first-grade public school students. (See Gómez-Bellengé & Thompson, 2005.) Keep in mind that the norms for the Observation Survey are based on a random sample of first-grade children and represent the performance of all first graders, not only struggling readers. The new norms are probably more representative of the range of literacy achievement found among first-grade children than the 1990–1991 norms were in their time. Even so, a comparison between the two time periods is revealing.

For brevity, we will examine the Text Reading Level task only. In 1990–1991, stanine 5 (average reading level) in fall corresponded to a text level of 2. In 2002–2003, stanine 5 corresponded to a raw score of 3–5. This is not such a big change. At mid-year, stanine 5 corresponded to text levels 6 and 7; today, it is 12 to 14. At year-end, the change is from text levels 9 to 12, to text level 20 (Pinnell, McCarrier, & Button, 1990; Gómez-Bellengé & Thompson, 2005). These are greater changes in overall achievement levels (by all

Table 1. The New U.S. Stanines

Stanines are a type of score based on the mean and standard deviation of scores on a task. They indicate how different from average a particular raw score is. The mean score belongs to stanine 5. In some cases, these stanines have been adjusted for the purpose of student selection. For utilizing the new stanines, Reading Recovery professionals are asked to use fall stanines from August to November, mid-year stanines from December to February, and year-end stanines from March to July.

Letter Identification

Purpose To find what letters a child knows and the preferred mode of identification.
Task Identify upper- and lower-case letters and print forms of "a" and "g".
Scoring Maximum score = 54.

Stanines for Letter Identification

Period	1	2	3	4	5	6	7	8	9
Fall	0-38	39-41	42-44	45-48	49-51	52-54	—	—	—
Mid-Year	0-47	48	49-50	51	52-53	54	—	—	—
Year-End	0-49	50	51	52	53	54	—	—	—

Concepts about Print (CAP)

Purpose To find what a child has learned about how spoken language is put into print.
Task Perform a variety of tasks during book reading by the teacher.
Scoring Maximum score = 24.

Stanines for Concepts about Print

Period	1	2	3	4	5	6	7	8	9
Fall	0-7	8-9	10-11	12-13	14-15	16-17	18-19	20	21-24
Mid-Year	0-13	14	15-16	17	18-19	20	21-22	23	24
Year-End	0-15	16-17	18	19	20-21	22	23	24	—

Ohio Word Test (OWT)

Purpose To find if a child is developing a personal resource of reading vocabulary.
Task Read a list of high-frequency words.
Scoring Maximum score = 20.

Stanines for Ohio Word Test

Period	1	2	3	4	5	6	7	8	9
Fall	0*	1*	2*	3-5	6-8	9-11	12-14	15-17	18-20
Mid-Year	0-5	6-8	9-10	11-13	14-15	16-18	19-20	—	—
Year-End	0-13	14-15	16	17	18-19	20	—	—	—

*** The raw scores in this stanine do not correspond to the mean and standard deviation for this task. They have been adjusted for the purpose of student selection.**

children, not just Reading Recovery children).

Compared to basal equivalents, the old spring stanines corresponded to basal level 1–1 (first semester of first grade). The new spring stanines correspond to basal level 2–2 (second semester of second grade). Even with the limitations of such a comparison, these data suggest the average reading level of U.S. first-grade children is probably higher today than it was a decade or so ago.

Our definition of average has also changed. The low end of stanine 4 is now text level 16. Any first-grade child reading below text level 16 at year-end is now below average, if one is to accept this classification. Children reading below level 16 in spring of first grade are among the lowest 23% of readers.

The new norms and stanines give us a lot of food for thought. Student selection procedures have changed somewhat. The reading performance of most first-grade children is higher than it was 10 or so years ago. These higher standards for early reading are reflected in the new norms. But what does this mean for our lowest readers?

The national average text reading level in spring for children who successfully discontinued their series of Reading Recovery lessons was 19.4 in 2004, compared to 20 for the random sample, meaning these children were reading at average levels following the intervention. This is good news.

However, a sobering thought is that while students are entering first grade at higher reading levels than in past years, struggling readers are no better prepared for first grade than in past years. In other words, for the lowest readers, the performance gap in fall, before Reading Recovery, is getting larger (Gómez-Bellengé, Rodgers, & Schulz, 2005). The need for Reading Recovery is as urgent as it has always been.

References

Clay, M. M. (1993). *An observation survey of early literacy achievement.* Portsmouth, NH: Heinemann.

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

College Board (2005). New SAT® for the Press: Glossary of Terms. collegeboard.com [Online]. Available: http://www.collegeboard.com/about/news_info/sat/glossary.html

Fried, M. D., Briggs, C. M., Konstantellou, E., & Rosser, M. (2005). *Procedures for administering leveled text passages and stanines for the observation survey. Working document for 2005–2006.* Columbus, OH: North American Trainers Group.

Gómez-Bellengé, F. X., Rodgers, E. M., & Schulz, M. (2005). *Reading Recovery and Descubriendo la Lectura national report 2003–2004.* Columbus: The Ohio State University.

Gómez-Bellengé, F. X. , & Thompson, J. R. (2005). *U.S. norms for tasks of an observation survey of early literacy achievement* (Rep. No. NDEC 2005–02). Columbus: The Ohio State University, National Data Evaluation Center. http://www.ndec.us

Hoover, H. D., Dunbar, S. B., Frisbie, D. A., Oberley, K. R., Ordman, V. L., Naylor, R. J., et al. (2003). *The IOWA tests. Interpretive guide for teachers and counselors. Forms A and B. Levels 5–8.* Itasca, IL: Riverside Publishing.

Paris, S. G.,& Hoffman, J. V. (2004). Reading assessments in kindergarten through third grade: Findings from the Center for the Improvement of Early Reading Achievement. *Elementary School Journal, 105,* 199–217.

Pinnell, G. S., McCarrier, A., & Button, K. (1990). *Constructing literacy in urban kindergartens: Progress report on the kindergarten early literacy project.* Columbus: The Ohio State University.

Table 1. ***continued***

Writing Vocabulary (WV)

Purpose: To find if a child is building a personal resource of words that can be written.

Task: Write all known words in 10 minutes.

Scoring: Count of words in a 10-minute time limit.

Stanines for Writing Vocabulary

Period	1	2	3	4	5	6	7	8	9
Fall	0-1*	2*	3-7	8-13	14-19	20-24	25-30	31-36	37+
Mid-Year	0-12	13-20	21-28	29-35	36-43	44-51	52-58	59-66	67+
Year-End	0-22	23-31	32-40	41-49	50-58	59-67	68-76	77-85	86+

*** The raw scores in this stanine do not correspond to the mean and standard deviation for this task. They have been adjusted for the purpose of student selection.**

Hearing and Recording Sounds in Words (HRSIW)

Purpose: To assess phonemic awareness by determining how well a child represents the sounds of letters and clusters of letters in graphic form.

Task: Write a dictated sentence, with credit for sounds correctly represented.

Scoring: Maximum score = 37.

Stanines for Hearing and Recording Sounds in Words

Period	1	2	3	4	5	6	7	8	9
Fall	0-5	6-10	11-15	16-20	21-25	26-30	31-35	36-37	—
Mid-Year	0-24	25-27	28-29	30-32	33-34	35-37	—	—	—
Year-End	0-29	30-31	32	33-34	35-36	37	—	—	—

Text Reading Level (TRL)

Purpose: To determine an appropriate level of text difficulty and to record, using a running record, what a child does when reading continuous text.

Task: Read texts representing a gradient of difficulty until the highest text level with 90% or better accuracy is determined, with teacher recording behaviors during the oral reading.

Scoring: Maximum score = 30.

Stanines for Text Reading Level

Period	1	2	3	4	5	6	7	8	9
Fall	0*[1]	0*[2]	1*	2*	3-5	6-8	9-10	12	14-30
Mid-Year	0*	1-2*	3-6	7-10	12-14	16-18	20-22	24-26	28-30
Year-End	0-6	7-10	12-14	16-18	20	22-24	26-28	30	—

*** The raw scores in this stanine do not correspond to the mean and standard deviation for this task. They have been adjusted for the purpose of student selection.**

[1] A child who does not read the Level 1 text with 90% accuracy *even after it has first been read entirely by the teacher* is assigned a score of zero for data collection. A child who has an opportunity to read his own dictated text as written and reread by the teacher is also assigned a score of zero. Use stanine 1 for student selection in fall.

[2] A child who does read the Level 1 text with 90% accuracy *only after it has first been read entirely by the teacher* is assigned a score of zero for data collection. Use stanine 2 for student selection in fall.