

CHAPTER 9

TOWARD A UNIFORM ENTRY/EXIT ACADEMIC ASSESSMENT

9.1 Introduction

Over the past few decades, the field of education has become increasingly interested in standardized testing that identifies, measures, and compares outcomes at all levels, including national, state, district, school, teacher, and ultimately, the level of the individual student. A number of concerns drive this interest, including the relatively low performance of American students (as compared to other industrialized nations), a general public perception of unacceptably low levels of educational achievement, and the attendant criminogenic problems that arise from a poorly educated population. With the recent implementation of the federal No Child Left Behind Act (NCLB), every state is now required to develop and implement consistent outcome measures, including measures of academic gains among students that can be connected to a specific school or program. Many states have already designed, and even implemented, educational measures that have anticipated the intent of NCLB, including Florida where the Juvenile Justice Educational Enhancement Program (JJEPP) is charged with research and evaluation of the educational component of Florida's juvenile justice system.

To that end, JJEPP uses diverse measures, which include annual quality assurance (QA) reviews, Florida Department of Education (FLDOE) Survey Five data, teacher certification information, and longitudinal study of community reintegration results. These measures, though separate, triangulate on an underlying common factor: the quality of the educational opportunities afforded students by their respective Department of Juvenile Justice (DJJ) program. High-quality educational opportunities, if taken advantage of, should lead to academic achievement and successful community reintegration. While QA serves as an indicator of the quality of educational opportunity, investigating its relationship to student performance is not an easy task. Comparing individual student academic gains achieved while in programs using different test instruments is often impossible because of incompatible scoring systems and different norm groups. Additionally, confounding variables may exist, such as DJJ school characteristics (e.g., provider type and security level) and student characteristics (e.g., socioeconomic status, age, and gender) that can obscure the relationship between quality educational opportunities and academic gains. Nevertheless, FLDOE must develop a method of assessing academic gains within DJJ schools in order to comply with the requirements of NCLB.

This chapter reviews the various student assessments used to measure academic gains in DJJ programs in Florida, and is comprised of five subsequent sections. Section 9.2 describes the Florida Comprehensive Assessment Test (FCAT), a well known standardized test given to every student in Florida from grades three through 10, and explains why the FCAT is not an

effective tool for measuring student academic gains in DJJ programs. Section 9.3 outlines the current status of entry/exit assessment in DJJ schools. Section 9.4 discusses the advantages and disadvantages of the various assessments in common use among programs. Section 9.5 describes the need for a common assessment. Section 9.6 provides a summary and discussion of future implications for policy makers and educators in Florida regarding uniform entry/exit academic assessment.

9.2 The FCAT

Recognizing the need for universal standards and accountability throughout the state of Florida, educators began development of a set of content and skill standards in the 1990s that would identify what students should know at each grade level. The results of this effort were known as the Sunshine State Standards. These standards were created to ensure that teachers were providing a baseline level of education, thereby creating a universal curriculum that would adequately meet the educational needs of Florida's youths and develop a consistent mechanism for student, teacher, school, district, and state monitoring. Florida's educators and political leaders recognized a need to provide a basic, standard education with a universal set of skills and content knowledge to students who were at the same grade level, regardless of their location in Florida, while still allowing quality teachers the creative latitude to teach to the Florida Sunshine State Standards (FSSS). Thus, the convergent needs of tracking student performance to ensure consistency and identify deficiencies, while also holding teachers, schools, districts, and ultimately the state itself accountable, led to the development of a test based on the FSSS -- the FCAT.

Implementation of the FCAT began in 1997, replacing an earlier, limited statewide assessment known as the High School Competency Test (HSCT). The FCAT was expanded gradually, each year including additional grade levels and subject areas. FCAT creators field-tested and evaluated each item on the FCAT to ensure that the test was fair, appropriate, nonbiased, and matched the FSSS.

The current FCAT is actually comprised of two distinct tests. The first of these is the criterion-referenced exam, which tests students on the content and skills as set forth in the FSSS in reading, writing, science, and mathematics. The second test is the Stanford Achievement Test ([SAT], though not to be confused with the college admissions exam with the same acronym). The SAT is a nationally norm-referenced test that provides an indication of how well Florida's students perform compared to their peers across the nation.

Current FCAT tests are administered as illustrated in Table 9.2-1.

Table 9.2-1: Administration of the FCAT for 2003-2004

| <i>Grade</i> | <i>Reading</i> | <i>Writing</i> | <i>Math</i> | <i>Science</i> |
|--------------|----------------|----------------|-------------|----------------|
| 3 | ✓ | | ✓ | |
| 4 | ✓ | ✓ | ✓ | |
| 5 | ✓ | | ✓ | ✓ |
| 6 | ✓ | | ✓ | |
| 7 | ✓ | | ✓ | |
| 8 | ✓ | ✓ | ✓ | ✓ |
| 9 | ✓ | | ✓ | |
| 10 | ✓ | ✓ | ✓ | ✓ |

Note. Passing the FCAT in the 3rd and 10th grades is necessary for promotion/graduation.

In addition to the required FCAT testing in grades three through 10, third graders must “pass” the FCAT; that is, they must attain an acceptable score in reading to be promoted to the fourth grade and beyond. Furthermore, state law requires high school seniors to pass the 10th grade FCAT before receiving a standard diploma.

Score levels on the FCAT range from one to five, with five being the highest as described in Table 9.2-2. Scoring for the writing assessment is handled differently and not discussed here.

Table 9.2-2: FCAT Score Descriptions

| <i>Score Level</i> | <i>Description</i> |
|--------------------|---|
| 5 | Performance at this level indicates the highest achievement. A level 5 student has success with the most challenging content of the <i>Sunshine State Standards</i> and correctly answers most of the test questions. |
| 4 | Performance at this level indicates that the student has success with the content of the <i>Sunshine State Standards</i> and correctly answers many of the most challenging test questions. |
| 3 | Performance at this level indicates that the student has partial success with the content of the <i>Sunshine State Standards</i> and correctly answers many of the test questions but is generally less successful with the most challenging questions. |
| 2 | Performance at this level indicates that the student has limited success with the challenging content of the <i>Sunshine State Standards</i> . |
| 1 | Performance at this level indicates that the student has little success with the challenging content of the <i>Sunshine State Standards</i> . |

Note. Descriptions provided by DOE at fcats.fdoe.org.

Currently, juvenile justice educational programs in Florida participate in administration of the FCAT along with their public school counterparts. The FCAT is given on the same dates to juvenile justice students, who must take the test regardless of their status within the juvenile justice facility or educational program. This means that students who have recently arrived from situations where they might have attended little or no school and those who have been in a juvenile facility for more than a year must all take the FCAT at exactly the same time. It is immediately clear that these unique operating conditions limit the utility of the FCAT in terms of assessing the academic gains of DJJ students. The FCAT was designed for 'normal' circumstances in which there is limited student mobility; teachers, schools, and home lives remain more or less consistent; and critical life events are the exception rather than the rule. In the juvenile justice population, students and their families are highly mobile and far less likely than their nondelinquent counterparts to remain with the same teachers and schools, creating a lack of continuity of education and instruction.

The FCAT cannot effectively serve as an accountability tool for juvenile justice teachers or programs because few students remain for extended time periods in a single program with the same teachers. Instead, FCAT scores among juvenile justice students may reflect more accurately on whatever school the student attended prior to entering the juvenile justice educational program. Although, as discussed later in this chapter, there are problems with any entry assessment test that is administered too soon after a student is admitted to a DJJ program, when FCAT exam dates fall near the DJJ program entry date of a student, the results may be a useful tool to assess needs. When FCAT exam dates fall close to the exit date of a student who has been in the DJJ program since before the previous year's administration of the FCAT, the results may even be able to show academic gains while in the program. Given the high mobility of juvenile justice students, however, test results may be of no use in determining program influence on any observed academic gains.

Finally, because of the significantly higher percentage of exceptional student education (ESE) students in juvenile justice facilities (see Chapter 3), accommodations on the FCAT are an important issue in juvenile justice educational programs. Not only is there a greater ESE population, but also ESE teachers and the training and assistance in providing those important testing accommodations may be limited. Additionally, many juvenile justice students who would qualify as ESE students simply have not been identified and may, therefore, not be receiving the accommodations they need to successfully take the test, in which case the results for these students might be compromised.

In juvenile justice education, the FCAT serves at least two useful functions. First, juvenile justice FCAT scores may be compared to the FCAT scores of their regular school counterparts to determine general educational deficiencies and needs in the juvenile justice population. Second, as required by legislative statute, the FCAT remains a requirement for obtaining a high school diploma and must be administered to afford juvenile justice students the opportunity to advance to the next grade level. Nevertheless, despite these worthwhile uses of the FCAT for DJJ students, the FCAT is inappropriate for measuring academic gains among this highly mobile population. Currently, other academic assessment tools are available and discussed in the following section.

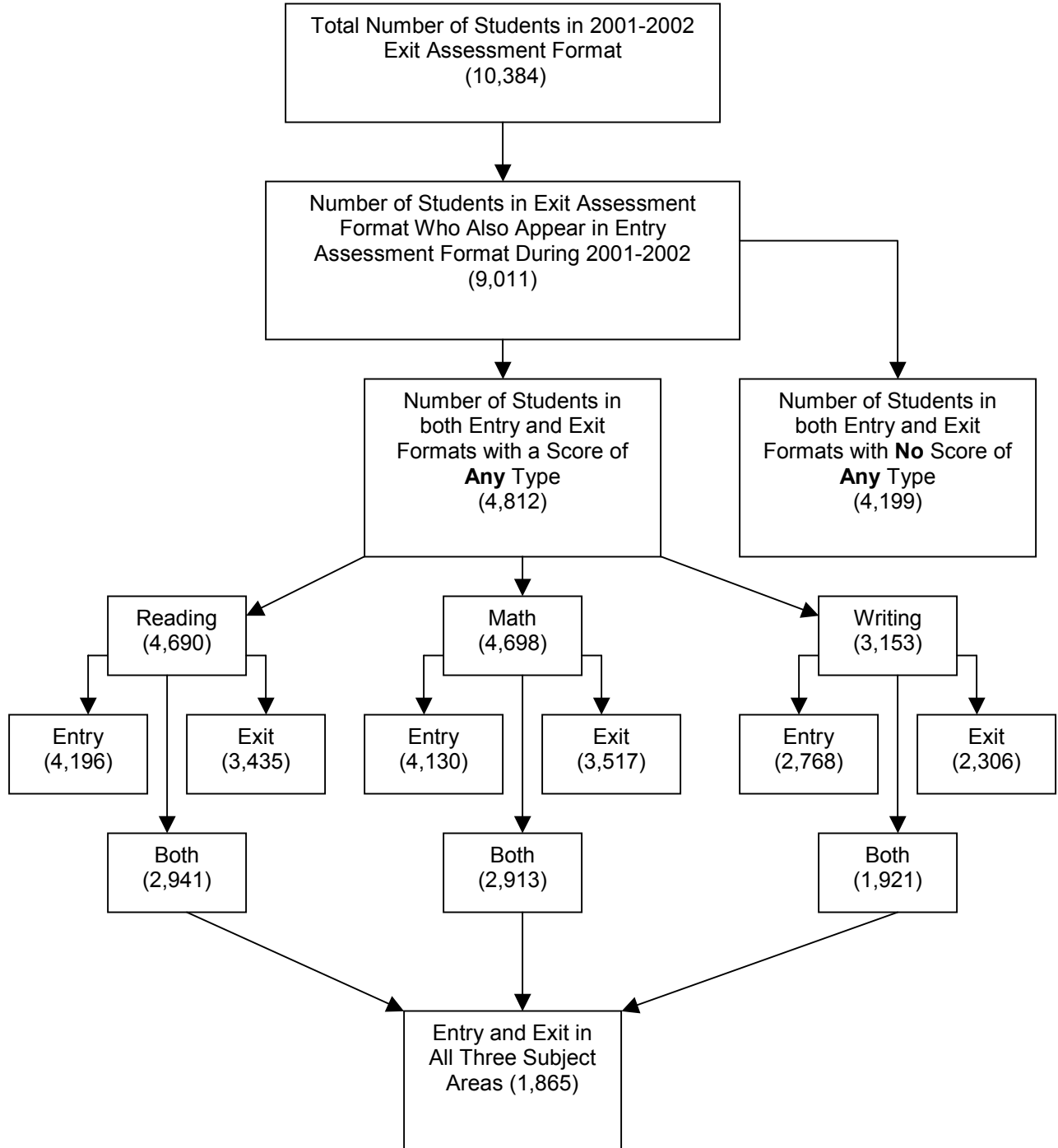
9.3 Other Academic Assessment Tools Currently in Use

Section 1003.51, F.S., requires the FLDOE, in partnership with DJJ, district school boards, and private providers, to develop procedures for the administration of entry and exit academic assessments in DJJ facilities. Rule 6A-6.05281, FAC, further clarifies this requirement to include academic entry and exit assessments that measure student performance in the areas of reading, writing, and math. Additionally, the rule requires all residential commitment and day treatment programs to report the assessment test results of students to the local school district management information system (MIS) and include them in FLDOE Survey 5 data. This reporting process began in 2002.

Also in 2002, the FLDOE developed and disseminated the technical assistance paper (TAP), *A Guide to Test Instruments for Entry and Exit Assessment in Florida Department of Juvenile Justice Educational Programs*. The TAP provides information on 32 different tests that have been approved for use as entry and exit assessments in juvenile justice educational programs. The TAP also describes the process for reporting student test scores to DOE. The approved assessments are scored using several different rubrics. The TAP also includes a section on the proper administration of academic assessment tests.

Figure 9.3-1 shows the breakdown of students present in the entry and exist assessment Survey Five formats for 2001-2002.

Figure 9.3-1: Entry/Exit Assessment Flow Chart



The exit assessment format should serve as an indication of the number of youths, released from juvenile justice programs, who received educational services during their stay. Because some programs have lengths of stay longer than one year, and others may admit students toward the end of one school year and not release them until after the start of the next school year, the subset of those youths who also appear in the entry assessment format is a reasonable measure of all students who **should** have been given both an entry and exit assessment during the 2001-2002 school year.

Among the 9,011 students who appeared in both the entry and exit assessment formats for Survey 5 in 2001-2002, only 53% had any assessment information. The writing assessment was missing much more often than math or reading. In the end, only 21% of students had both entry and exit assessment scores for all three subject areas.

Table 9.3-1 shows the number and percentage of DJJ schools using each assessment. Although the TAP lists 32 FLDOE-approved assessments, only half of them have been used in DJJ schools. There were 179 DJJ schools that reported information in the FLDOE exit assessment format in 2001-2002.

Table 9.3-1: Assessments Used by DJJ Schools in 2001- 2002

| <i>Reading</i> | | | <i>Math</i> | | | <i>Writing</i> | | |
|---|---------------------------|----------|---|---------------------------|----------|---|---------------------------|----------|
| <i>Test</i> | <i>Number of Programs</i> | <i>%</i> | <i>Test</i> | <i>Number of Programs</i> | <i>%</i> | <i>Test</i> | <i>Number of Programs</i> | <i>%</i> |
| Tests of Adult Basic Education 7&8 | 164 | 92.1 | Tests of Adult Basic Education 7&8 | 164 | 92.1 | Tests of Adult Basic Education 7&8 | 164 | 92.1 |
| Standard Test for Assessment of Reading | 119 | 66.9 | Standard Test for Assessment of Reading | 127 | 71.3 | Test of Written Language-3 | 80 | 44.9 |
| Wide Range Achievement Test 3 | 119 | 66.9 | Wide Range Achievement Test 3 | 121 | 68.0 | Woodcock Johnson Tests of Achievement-III | 60 | 33.7 |
| New Century | 110 | 61.8 | New Century | 110 | 61.8 | Mini-Battery of Achievement | 59 | 33.1 |
| Woodcock Johnson Tests of Achievement-III | 72 | 40.4 | Woodcock Johnson Tests of Achievement-III | 73 | 41.0 | Hammill Multiability Achievement Test | 46 | 25.8 |
| Scholastic Reading Inventory | 49 | 27.5 | Kaufman Test of Educational Achievement | 54 | 30.3 | Basic Academic Skills Individual Screener | 4 | 2.2 |
| Slosson Oral Reading Test | 33 | 18.5 | Key Math Revised | 39 | 21.9 | Wide Range Achievement Test 3 | 4 | 2.2 |
| Woodcock Reading Mastery Test-Revised | 32 | 18.0 | Mini-Battery of Achievement | 27 | 15.2 | | | |
| Kaufman Test of Educational Achievement | 30 | 16.9 | Hammill Multiability Achievement Test | 12 | 6.7 | | | |
| Mini-Battery of Achievement | 27 | 15.2 | | | | | | |
| Hammill Multiability Achievement Test | 12 | 6.7 | | | | | | |
| Total* | 195 | | Total* | 190 | | Total* | 115 | |

Note. There were 178 DJJ schools that submitted data to the exit assessment format during Survey 5 in 2001-2002. There is some overlap in the number of assessments used because DJJ schools often used more than one type of assessment within a subject area. Additionally, where the TABE was administered to students younger than 16, a program might have reassessed the student during the year with an age-appropriate instrument.

As indicated in Table 9.3-1, during the 2001-2002 school year, the most commonly used tests for reading and math were the Tests of Adult Basic Education (TABE 7&8), the Wide Range Achievement Test 3rd Edition (WRAT-3), the Standardized Test for Assessment of Reading (STAR), the New Century Education (New Century), and the Woodcock-Johnson 3rd Edition (WJ-III). Although the Scholastic Reading Inventory (SRI) is used more frequently to assess reading than the New Century, the SRI does not have a math component. Among those schools that tested students in reading and math during 2001-2002, more than 80% used at least one of the aforementioned assessments. Similarly, DJJ schools predominantly used the TABE, the MBA, and the WJ-III to assess writing.

Since each DJJ school serves a different number of students for varying lengths of time, it also is useful to examine how many students have been tested using the assessments. Table 9.3-2 shows the percentage of students who took each test when they entered their DJJ school.

Table 9.3-2: Assessments used for Students Who Exited DJJ Schools in 2001-2002

| <i>Reading</i> | | | <i>Math</i> | | | <i>Writing</i> | | |
|---|---------------------------|-------------|---|---------------------------|-------------|---|---------------------------|------------|
| <i>Test</i> | <i>Number of Students</i> | <i>%</i> | <i>Test</i> | <i>Number of Students</i> | <i>%</i> | <i>Test</i> | <i>Number of Students</i> | <i>%</i> |
| Test of Adult Basic Education 7&8 | 2,268 | 21.8 | Test of Adult Basic Education 7&8 | 2,259 | 21.8 | Test of Adult Basic Education 7&8 | 2,484 | 23.9 |
| Wide Range Achievement Test 3 | 565 | 5.4 | Standard Test for Assessment of Reading | 705 | 6.8 | Test of Written Language-3 | 176 | 1.7 |
| Standard Test for Assessment of Reading | 561 | 5.4 | Wide Range Achievement Test 3 | 592 | 5.7 | Woodcock Johnson Tests of Achievement-III | 112 | 1.1 |
| New Century | 462 | 4.4 | New Century | 462 | 4.4 | Mini-Battery of Achievement | 102 | 1.0 |
| Woodcock Johnson Tests of Achievement-III | 179 | 1.7 | Woodcock Johnson Tests of Achievement-III | 180 | 1.7 | Hammill Multiability Achievement Test | 71 | 0.7 |
| Scholastic Reading Inventory | 74 | 0.7 | Kaufman Test of Educational Achievement | 86 | 0.8 | Basic Academic Skills Individual Screener | 4 | 0.0 |
| Slosson Oral Reading Test | 44 | 0.4 | Key Math Revised | 55 | 0.5 | Wide Range Achievement Test 3 | 4 | 0.0 |
| Kaufman Test of Educational Achievement | 42 | 0.4 | Mini-Battery of Achievement | 35 | 0.3 | | | |
| Woodcock Reading Mastery Test-Revised | 41 | 0.4 | Hammill Multiability Achievement Test | 12 | 0.1 | | | |
| Mini-Battery of Achievement | 35 | 0.3 | | | | | | |
| Hammill Multiability Achievement Test | 12 | 0.1 | | | | | | |
| No Test | 6,101 | 58.8 | No Test | 5,998 | 57.8 | No Test | 7,431 | 71.6 |
| Total Students | 10,384 | 99.8 | Total Students | 10,384 | 99.9 | Total Students | 10,384 | 100 |

Note. Percentages may not total 100% due to rounding.

Of the students tested in 2001-2002 for reading and math, more than 90% of students were tested using the TABE 7&8, the STAR, the WRAT-3, the New Century, and/or the WJ-III. Similarly, to assess writing, the TABE 7&8, the MBA and the WJ-III were used for the vast majority of students.

9.4 Advantages and Disadvantages of the Different Assessments

Despite the relatively small number of different assessments in use, comparisons among students tested remains problematic for several reasons. First, the target groups of each assessment vary. For instance, the most widely used test, the TABE, is designed for students who are at least 16 years old; the STAR and the New Century are designed for students in grades one through 12; the WRAT-3 is designed for ages five-to-adult, and the WJ-III for ages two-to-adult. Second, there are variations in the subject areas tested. Some tests cover only one or two subject areas even though the FLDOE requires DJJ schools to assess students in reading, math, and writing. For instance, the STAR and the WRAT-3 do not include a writing component. Finally, even when tests assess the same academic subject, the content areas may vary from test to test. Some reading assessments, for example, cover only reading comprehension (e.g., the STAR), reading comprehension and spelling (e.g., the WRAT-3), or reading fluency and spelling (e.g., the WJ-III). Most math assessments are comprehensive, but some tests (e.g., WRAT-3) only assess arithmetic ability.

There are also differences in testing methods among the different assessments. Traditional testing methods include the use of paper and pencil tests; however, newer tests (such as the STAR and WJ-III) may be administered using a computer. According to the STAR manual, a computer-based test using “adaptive” testing methods to adjust its difficulty level to test taker’s responses, may produce more reliable test results.

For the purpose of inter-test comparison, of particular interest is the scoring system each assessments employs. One method of scoring involves percentile ranks (PR), where a student’s score is ranked against that of other respondents and reported at the percentile representing the proportion of respondents who scored lower on the test than did the student. Another method uses the normal curve equivalent (NCE) and assigns the student a score that corresponds to a point on the normal curve that can be expressed in standard deviations above or below the mean. A standard nine or STANINE scale scoring system assigns respondents a score on a nine-point scale such that the mean is five, and standard deviation is two. Finally, a grade equivalent (GE) score assigns the school grade (K-13) to which the student’s responses correspond. A decimal is sometimes used to denote months in that grade as a way to add variability to scores among students performing at the same grade level who have different ability levels.

It is well known that testing circumstances influence student performance (Campbell & Stanley, 1978). Testing time, place, and other circumstances differ in each DJJ school. Of interest is who administers a test and, furthermore, how and when it is administered. The TAP and testing manuals require either an educational diagnostician or a student service

professional as a qualifying tester; however, many programs do not have such qualified individuals. Furthermore, literature on assessment testing has documented that students should not be assessed immediately upon entry into a new school environment. Nonetheless, section 1003.51, F.S., requires that all DJJ students be assessed within five days of entry.

Table 9.4-1 outlines the scoring system, normed group, age range, and strengths and weaknesses of each of the major tests in current use in Florida DJJ schools.

Table 9.4-1: Characteristics of Reading Test Instruments Used by DJJ Schools

| <i>Test</i> | <i>Scoring System(s)</i> | <i>Normed Group(s)</i> | <i>Age Appropriate Range</i> | <i>Advantages</i> | <i>Disadvantages</i> |
|-------------|---|---|------------------------------|--|--|
| TABE | GE, Percentiles, Stanine, Scale Score | (1) Adult basic education enrollees; (2) Vocational/technical school enrollees; (3) Adult/juvenile offenders; (4) College students | 16 and up | Widely used; covers reading, math, and writing | Inappropriate for 40% of population who are younger than 16. |
| WRAT | GE, Percentile, Stanine, NCE, Raw Score, Absolute Score, Standard Score | Age cohorts on the national level | Ages five-to-adult | Appropriate for all ages of DJJ students; covers reading, math, and writing | Measures only arithmetic skills in math and word reading ability in reading. |
| STAR | GE, Percentile, NCE, Scaled Score, Instructional Reading Level | The same grade peer on the national level | Grades 1 to 12 | Appropriate for all ages of DJJ students; employs adaptive testing format | Does not contain a writing or language arts component |
| New Century | GE | Grade peers on the national level | Grades 1 to 12 | Appropriate for all ages of DJJ students; | Only uses GE and reports only as 1 st or 2 nd semester instead of the full decimal |
| WJ III | GE, Age Equivalents, Percentile, Standard Score | Age cohorts and Grade peers on the national level | Ages two to adult | Appropriate for all ages of DJJ students; employs adaptive testing format; covers reading, math, and writing | Difficult to score. Requires tester to hold a master's degree. |

Setting aside the inherent difficulties that stem from attempting to standardize the scoring systems employed by each of these tests, it is clearly evident that each has both strengths and weaknesses. Even so, several common problems with the assessments themselves emerge, making some less attractive as candidates for system-wide implementation than others.

9.5 The Need for A Common Assessment

One of JJEEP's major research initiatives is to determine whether quality education leads to better academic achievement. The QA review process operationalizes quality education by uniformly measuring features such as percentage and type of qualified teachers, class size, support services for students, and individual attention. While it is relatively easy to compare and contrast the quality of education among DJJ schools because of these more or less

standardized measures, determining the relationship between program educational quality and academic achievement of participants is considerably more difficult. This is due to the lack of a uniform assessment in each academic subject area. A uniform, standardized assessment battery designed for the juvenile justice student is essential for establishing the strength of this relationship and for determining under what programmatic conditions certain sub-populations are most likely to succeed academically.

A large volume of research has attempted to compare test instruments that measure the academic abilities of students (Bray & Estes, 1975; Jenkins & Pany, 1978; Jones & Armitage, 1984; McCabe, Marglis, & Barenbaum, 2001; Prewett & McCafery, 1993; Sabatini, Venezky, & Bristow, 1995). The focus of these studies, however, has been correlation *between* instruments, while JJEEP's interest is to directly compare individual achievement over time. High correlation between tests is, therefore, of limited utility, since a high correlation merely indicates that scores among tests vary in the same direction. For instance, a test that systematically overestimates academic scores can be highly correlated with another test that systematically underestimates scores, as long as both tests vary in the same direction with regard to the underlying population being measured.

It is also often difficult, and in many cases impossible, to convert scores assigned using one system to a different measurement scale without distorting variability. Even when scales are the same across tests, the norm groups on which they are based may be different. For instance, the percentile rank on the STAR represents "how an individual student's performance compares to that of his or her same-grade peers on the national level" (STAR manual, p. 48). The norm group on the WRAT-3 is age-peers, however, and norm reference groups on the TABE are drawn from four different cohorts: adult basic education enrollees, vocational/technical school enrollees, adult/juvenile offenders, and college students.

Grade equivalency (GE) scores represent the lowest common denominator to which any of the other scales can be converted because they are normed against peers nationwide, share a common measurement scale, and because some tests (e.g., New Century) do not report any score *except* GE (e.g., Jones & Armitage, 1984). Important caveats should, nevertheless, be emphasized when comparing GE scores across tests. First, the meaning of GE may vary in each test, although test providers attempt to make it compatible across tests. Sampling methods and areas of testing can be sources of such discrepancies. Second, reported GE scores are simply estimates with different reliability and confidence intervals (School Renaissance Institute, 2000). Therefore, a one or two GE score disparity may simply be an artifact of chance and not a measure of true variability, particularly when the scores are obtained from different tests. For example, Jones and Armitage (1984) found that when Navy recruits took three reading tests (TABE, Nelson-Denny, and Gates-MacGinitie), their average GE scores varied significantly from test to test. The 95% confidence interval fluctuated from 8.88 to 11.45. Third, GE scores may differ across tests due to differences in testing formats. According to the STAR Norms/Technical Manual (2003), computer-adaptive testing formats such as STAR provide more consistently accurate scores than do classical non-adaptive tests. "GE obtained using classical test instruments are less accurate when a student's grade placement and GE score differ markedly," and it is "not uncommon for a fourth grade student to obtain a GE score of 8.9" when the student answers nearly all

items correctly (p. 44). Finally, and perhaps most importantly, the GE scale itself does not contain enough variability to be useful for computing academic *gains* among students who are in DJJ programs for short periods of time, which are often less than a semester. For these reasons, the use and conversion to GE scale scores of differing assessments is not a viable option.

Therefore, in order to measure students' academic progress while in a DJJ program, a uniform scale of measurement with enough variability to detect academic gains over short periods of time is essential. This almost certainly requires abandoning the use of multiple tests in favor of a single assessment or, at the very least, a single assessment in each of the three academic areas where testing is required, normed appropriately for the population or sub-population being tested. This assessment must be reliable, valid, and designed for the target group being tested. It should measure students' mastery of FSSS skills and content to assess student strengths and deficiencies, and it should be available for administration as both an entry and exit test to provide both a measure of academic gains and to serve as an accountability tool for juvenile justice educational programs.

9.6 Summary Discussion

Despite the cautionary language above, JJEPP attempted to determine if any relationship between educational quality and academic gains could be detected using the data presently available. This preliminary attempt was limited to 2001-2002 reading and math assessment scores, since they are more widely used and reported than writing assessment scores. Despite the fact that accurate reporting is required by section 1003.51, F.S., and Rule 6A-6.05281, FAC, many reported scores were unusable. In most cases, DJJ schools did not report test names or scores. Even after limiting analyses to popular tests, such as the TABE 7&8, the WRAT-3, the New Century, the STAR, and WJ-III, less than 1/3 of students had usable data. The sample was further reduced due to apparent data entry or reporting errors, such as students who took both entry and exit tests on the same date, or received the exact same score at entry and exit. This indicates that schools may have simply reported the same score twice. Additionally, schools sometimes reported that students had lengths of stay that were zero days or even a negative number of days, indicating that either the entry date or the exit date (or both) were incorrect. In the end, fewer than 1,800 cases could be used for analysis. Therefore, in addition to problems with the disparate scoring systems, data entry and reporting problems must also be addressed before assessment information can be linked to educational quality.

A common academic assessment that addresses the NCLB and FSSS target areas is desperately needed in Florida for the delinquent population. The current practice of using any of 32 approved instruments does not allow for meaningful or accurate comparisons across programs or with non-delinquent peers. To this end, JJEPP has identified several key elements that any such assessment battery must contain to be useful when attempting to link educational quality with academic performance outcomes. At a minimum, the test must:

- be normed using the complete age range of students who are to take it
- report scores using percentile rankings against those norms

- address all relevant FSSS and NCLB subject areas
- have demonstrated internal and external reliability and validity

In addition, JJEEP has identified some effective testing conditions and procedures for students in juvenile justice facilities. The tests should:

- be administered as near as possible to student entry and exit from the program while still maintaining validity
- be administered in an environment conducive to maximizing student performance that is free from unnecessary distraction
- be consistently and accurately entered into FLDOE Survey Five data submissions in a timely manner.

The technical assistance paper (TAP), *A Guide to Test Instruments for Entry and Exit Assessment in Florida Department of Juvenile Justice Educational Programs* should be revised to specify testing time and place and the necessary qualifications of those who administer the test.