Running head: Review of Scantron Performance Series

Review of Scantron Performance Series

Jim Davis

ESR - 505 / National Louis University

Review of Scantron Performance Series

Chicago Public Schools (CPS) is shifting its assessment strategy for elementary schools to include the Scantron Performance Series computer adaptive test (CPS Office of Performance, 2009) . The new Chief Executive Officer of CPS (as of March, 2009), Ron Huberman, has introduced an aggressive program of "performance management" in the evaluation of school and teacher performance. This includes the use of "value-added" measures, which are intended to measure, among other things, student growth as indicators of performance (Meyers, 2009; Harris, 2010). Student growth, in turn, is measured in standardized test scores. Because the Scantron Performance Series provides an immediate, grade-independent scaled score, and will be administered three times a year, it will provide a regular and timely measure for the value-added metric of teachers and school. Because of its growing importance within Chicago public schools, it behooves all affected parties -- administrators, teachers, parents and students -- to know as much as possible about the new test regime. This paper reviews the Scantron Performance Series as an assessment instrument.

Overview

The Scantron Performance Series is a computer adaptive test to measure the proficiency level of students (Scantron, 2010a). The Performance Series assesses four areas: reading, mathematics, life sciences and scientific inquiry, and language arts (CPS is not using the language arts module). According to the publisher, the Performance Series has four primary uses: "more accurate student placement; diagnosis of instructional needs, including instructional adjustments; and measurement of student gains across reporting periods" (Scantron, 2010a). According to Scantron literature, the company develops its own reading passages and test items, based on an analysis of the skills required to meet various national and state standards (Scantron, 2010b). Test items are available for grades 2 through 12.

The Performance Series had its beginnings in the 1990s at the EdVISION Corporation , but the computer technology available at the time made the initial versions of the assessment difficult to administer. Cheaper and more powerful technology, including the Internet, expanded the feasibility of adaptive testing. EdVISION renamed their adaptive assessment product to Performance Series in 2001; and Scantron Corporation acquired the company in 2002. Scantron has continued to develop the product and market it since then (Scantron, 2010b). The current version of Performance Series is 6.1.4.

The Performance Series reports a customizable set of measures in its results, depending on the requirements of the purchaser. The core result number is a scaled, grade-independent performance score ranging from about 1300 to 3900. This number is used to measure student progress over time (Scantron recommends testing three times a year). In addition, similar scaled scores may be reported for each learning standard. The reading test will also generate a Lexile score and a reading rate. Scantron will also place the scaled score against national quartiles and percentiles to provide a national percentile ranking and a "grade level equivalency" (GLE) score. The standards scores can also be calibrated to report a percent probability of answers a student is likely to get correct of questions aligned to a particular standard for their grade level (called the Standard Item Pool (SIP) Score). It is interesting to note that CPS recently dropped the SIP Score and the GLE score from its reported results because "it has become evident that the GLE and SIP scores may be misinterpreted in ways that do not benefit students or teachers" (CPS, 2010).

Because the item bank is matched up with state and national learning standards, and the assessment is used nationally, the Performance Series can be used as a criterion-referenced assessment when results are interpreted against the state standards; and as a norm-referenced assessment when interpreted against the national pool of results (Scantron, 2010b). As a result, as noted by the publisher, Performance Series may serve multiple purposes. Because the assessment generates a grade-

independent scaled score, it may used to track student performance over several years. With the growing U. S. Department of Education emphasis on teacher evaluation based on student standardized test performance (see, e.g. Sawchuk, 2009; Illinois Government News Network, 2010; Medina, 2010), the Performance Series provides a metric for administrators to evaluate teachers and schools. Because student performance is mapped to state learning goals, the assessment may serve a diagnostic function, helping teachers identify student accomplishments and needs.

The Performance Series includes a variety of reports to help teachers and administrators make sense of assessment results. For the CPS configuration, student results are reported as a scaled score, and in a nationally-normed grade-level quartile of "below average", "low average", "high average" or "above average". Classroom results show each state performance objective covered, and the numbers of students who met and did not meet the objective. Links from the performance objectives area of the report lead to a bank of study guides and multiple choice quiz material. These materials can help teachers create custom study materials for students based on assessment results. The individual student profile report, which may be printed for sharing with the student or parents, includes a graph of student performance over test sessions, a scaled score for each learning strand, and a national percentile ranking. The reading report includes a Lexile score. Other scores may be available depending on the customer configuration. Aggregate reports are available for at the student, classroom, school and district level. Score, gain, and percentile reports are available to compare performance over test sessions.

The assessments themselves are administered on a computer with an Internet connection. Because it is an adaptive test, the testing time will vary per student. According to Scantron, the average number of test items a student sees is 50 (Scantron, 2010b). Some students may complete an assessment in as little as 15 minutes, while other students may need to work on an assessment for than

one hour. Students may pause the test, but the test must be completed within a two-week window. If a test is interrupted intentionally, or due to technical problems or the user accidentally exiting the test, the test will resume where the student left off. The Performance Series will spoil a student's test if the student answers questions too quickly, requiring the student to begin the test again. Administrators may also manually spoil tests.

Validity, Reliability, and Usability

For an assessment to be useful, it must be *valid, reliable*, and *usable* (Gronlund and Linn, 1995). An assessment is valid if it adequately and appropriately measures what it is intended to measure. An assessment is reliable if it yields consistent results over time. An assessment must also be practical to administer, and usable by the test subjects.

One measure of Performance Series validity is to compare results with other standardized instruments ("concurrent validity"). If another assessment, like the Illinois Standardized Achievement Test (ISAT), is deemed valid, and there is a strong positive correlation in results, then that suggests that Scantron is also valid. For reading, Scantron results have a .755 to .844 correlation to the Spring 2008 ISAT reading scores for grades 4 to 8. Math score correlations ranged from .749 to .823 (Scantron, 2010b). These numbers suggest that educators are seeing similar results in the two tests. There are other ways of assessing validity. Scantron literature describe two other dimensions of validity, "item validity" and "sampling validity." According to Scantron, item validity (how well to the test items assess skills they are designed to assess) begins with their item development process which is based on the company's standards database. The company then uses external consultants to review items. "Sampling validity" refers to the sample of test items selected on a test covering the entire subject area or strand being tested. This is of special importance for an adaptive assessment. Scantron uses the concept of "testlets" as a unit of measurement of content area subsets, and then compares performance

on testlets to ensure common results. The correlation among testlets is more than .65 (Scantron, 2010b). Due to their newness, similar data for the Language Arts and Science modules is not available, although test items for those modules go through the same development process as items for reading and math.

Scantron uses a "standard error of measurement" as one indicator of test reliability. This provides a range of scores that a student is likely to have if the test was repeated, Scantron's measure of reliability translates to a reliability coefficient of 0.91. Scantron argues that the adaptive test allows for a greater reliability than "fixed-form assessments" for students far above or below the grade level being tested, because it can draw on a flexible range of test items (Scantron, 2005). Another dimension of reliability relates to online assessments vs. paper and pencil test. The Scantron literature does not directly address this dimension:

> Scantron has not conducted any internal research to examine differences between administration of paper-and- pencil tests and online testing, since so much independent research has already been conducted in this area. A quick search of any educational research database should return many articles focusing on this topic. (Scantron, 2010b, p. 133)

Scantron then goes on to cite one source, as an example. But that source, a metadata analysis from 1993, only looked at tests on young adults and adults. Since Scantron sells its assessment system for grades 2 through 12, a comparison of paper vs. computer tests for younger students would be useful. Also, computer technology and interface design has gone through significant changes over the past 15-plus years, which might also affect reliability (and not necessarily in a negative way). In addition, it is plausible that a completely new test regime would yield skewed (i.e. unreliable) results, at least at first, as students become acclimated to the testing system.

Usability is difficult to measure quantitatively. There are a number of technical issues in administering the test, including equipment, the network connection, and database administration. The software itself has to be user-friendly. The test-taking experience should be neutral to the students. Based on personal observation, younger students may have difficulty with navigating the software at first, or be intimidated by the fact that a computer is challenging them with questions. The adaptive nature of the test means that students will be presented with material they do not know, and have not been exposed to, which can be demoralizing if they are used to tests that cover material that they have at least heard of or seen. Different kinds of cheating become possible for online tests, like using online dictionaries to look up words. Since the test must be administered on a computer, and for many schools, computers are located in labs, testing is done in a less-familiar setting that the regular classroom. Lab seating often is side-by-side, inviting wandering eyes, more distractions, and student interaction during testing. Reading long passages on the computer screen can be tiring for all students, and the adaptive nature of the reading test in particular frustrates students who are presented with multiple long passages. The online nature of the reading assessment precludes using reading strategies like highlighting. On the plus side, once the student database is set up, site corrdinators trained, and testing protocols are in place, the test is easy to administer, with test results available for analysis and action immediately.

Discussion

Scantron suffers from the limitations of any multiple-choice assessment. As far as standardized assessments go though, the Scantron Performance Series has a number of strengths. The criterion-referenced results can provide useful data for overall student assessment when combined with other data, both qualitative and quantitative. The grade-independent scaled score provides a convenient suggestion of student progress, though the potential for abuse by administrators is significant. The

software is easy to use, and the immediacy of the results is very useful. The challenge of online reading, especially with older cathode ray displays, is a significant issue, for the reading test primarily. As students become used to the testing regime, and test-taking protocols are developed, some of the limitations of the testing process will lessen or disappear. Other reviewers have noted that it is useful for high-level, long-term measurement of gains, but not for short-term or curriculum assessments (Henington, 2006). Another reviewer criticized the lack of science and language arts test result data, and considers the Performance Series as a "work in progress" (Morse, 2006). The same reviewer liked the relative ease of use, once the student database was set up.

If kept in its proper place, with proper protocols, Scantron can be a useful addition to educators' assessment toolbox.

References

Chicago Public Schools (2010). "Scantron Performance Series enhancements." Internal memo. Retrieved May 11, 2010 from https://research.cps.k12.il.us/export/sites/default/accountweb/Assessment/Scantron/NEWx_Score_Enhancements_in_Performance_Series.pdf

Chicago Public Schools Office of Performance (2009). Spring 2010 Elementary School Assessment Overview. Retrieved May 1, 2010 from https://research.cps.k12.il.us/cps/ accountweb/Assessment/Assessment_Overview_Presentations.html

Gronlund, N. and Linn, R. (1995). *Measurement and assessment in teaching*. Englewood Cliffs, NJ: Merrill.

Harris, R. (2010). "Helping principals, teachers the next wave of performance management." Catalyst notebook. Retrieved May 11, 2010 from http://www.catalyst-chicago.org/notebook/index.php/entry/533

Henington, C. (2006). Performance Series. Retrieved from Mental Measurements Yearbook database.

Meyers, J. (2009). "Huberman's performance SWAT team." Catalyst notebook. Retrieved May 11, 2010 from http://www.catalyst-chicago.org/notebook/index.php/entry/290/

Morse, D. (2006). Performance Series. Retrieved from Mental Measurements Yearbook database.

Sawchuk, S. (October 30, 2009). "Duncan calls for multiple measures in evaluation." *Education Week.* Retrieved May 15, 2010 from http://blogs.edweek.org/edweek/teacherbeat/2009/10/duncan_calls_for_multiple_meas.html

Scantron (2005). Above and beyond: Applying adaptive technology to diagnose student performance and progress. Scantron Corporation. Retrieved May 8, 2010 from http://www.scantron.com/downloads/Performance_Series_White_Paper.pdf

Scantron Performance Series (2010a). Scantron Corporation. Retrieved May 8, 2010 from http://www.scantron.com/performanceseries/

Scantron Performance Series technical report (2010b). Scantron Corporation. Retrieved May 11, 2010 from http://docs.achievementseries.com/docsPerformance/PerformanceSeriesTechnicalReport.pdf

Illinois government news network (2010). "Governor Quinn Signs Legislation to Improve Teacher, Principal Evaluations and Training." Press release. Retrieved May 15, 2010 from http://www.illinois.gov/PressReleases/ShowPressRelease.cfm?SubjectID=1&RecNum=8176

Medina, J. (May 10, 2010). Agreement will alter teacher evaluations. *New York Times*. Retrieved May 15, 2010 from http://www.nytimes.com/2010/05/11/nyregion/11teacher.html