

Validity and Reliability Statement Tests for Higher Standards

Reliability and validity of tests are twin pillars which support the entire testing enterprise. At its simplest, reliability is the consistency with which a test measures any attribute. In both the state-sponsored SOL tests and the various *TfHS* tests, this attribute is some type of academic proficiency. The validity of a test is how well it does indeed measure what it purports to measure. For both SOL tests and the *TfHS* tests the goal is to measure student proficiency on the competencies stated in the Virginia Standards of Learning. As the overall function of the *TfHS* products is to provide focus and feedback for instruction, it is essential that our tests be both valid and reliable! Below are some of the types of evidence we have collected or produced. We are grateful to the Chesapeake, Essex, Hopewell, and Orange school divisions for sharing some of their internal results with us.

TfHS Reliability. The evidence for the reliability for our tests presented here is based on the customary ***KR-20 internal consistency reliability estimates***. Each formula *KR-20* estimate is simply a statistical estimate of how each test question contributes to the overall score – averaged over all questions. It ranges from 0, no reliability, to 1, perfect reliability. Reliability estimates have been calculated for the current edition of our *Pre-Post Tests*. These were given in several grades and cover our five content areas (reading, writing, mathematics, science, and history and social science). Some example figures are presented in *Table 1* alongside some equivalent reliability estimates for the state-sponsored SOL Tests. The data were gathered from about 55 schools in several school divisions and were gathered in early spring of the year 2000.

<i>Grade</i>	<i>Reading</i>	<i>Writing</i>	<i>Mathematics</i>	<i>Science</i>	<i>History & Soc. Sc.</i>
5	.87 (.89)	.74 (.84)	.86 (.88)	.88 (.81)	.79 (.80)

Table 1. Internal Consistency Reliabilities (KR-20) for *TfHS Pre-Post Tests*. The numbers are adjusted to correspond to the lengths of the respective State SOL Tests using the Spearman/Brown formula (The numbers in parentheses are the equivalent KR-20 figures for the State SOL Tests.)

However, in comparing remember that the two sets of tests have quite different purposes. The state test is designed to make just one primary determination: “Is the student at or above the passing score in a subject area or course?” Any other use of the score, including diagnostic uses, is quite secondary to the primary purpose of the State SOL tests. The *TfHS Pre-Post Tests*, on the other hand, are mainly aimed at diagnosis. We ask, “On which of the SOL skills are the students proficient and where are their weak areas?” This is the information a teacher will need to craft instruction. (Note that *TfHS* has never set a “passing score” on these tests. That is not their purpose.) Although the *Pre-Post Tests* can determine how well a student or a group of students is doing overall, this use is usually secondary. The *TfHS Pre-Post Tests* are also longer than the State SOL Tests and each of them covers just one grade of Standards. (However, the *TfHS Simulation Tests* are designed to be as much like the State SOL Tests as possible.) As the *Pre-Post Tests* set out to measure many different proficiencies, the tendency to measure a single, overall traits may be slightly lower than the State SOL Tests. The reliabilities reported are highly appropriate for diagnostic tests.

TfHS Validity. The primary validity evidence appropriate to the various tests we publish is ***content validity***. That is, “Do the tests adequately reflect the content stated or implied by the Standards of Learning themselves?” By “content” we mean knowledge, understanding, skills, habits of mind, and so forth, contained in the standards. Content validity is established in the beginning by having our authors keep the standards directly in their view as we write, review, and revise test items. Each item is directed at measuring a specific, individual standard. We have always had teachers, administrators, and curriculum specialists carefully review all of our tests for content validity. Any item that appears upon review not to match its stated standard is removed from our test. There have been relatively few of these. Occasionally, we also receive comments from our users concerning the appropriateness of test content. All such comments are seriously reviewed and changes are made to our tests, where appropriate.

When the *TfHS* tests were first developed, we worked intensively and extensively with all teachers and with all students of one school division to ensure that each question and the tests as wholes measured the important content of the SOLs. As much as possible, we tried to follow development procedures similar to those being used to develop the State SOL Tests. We also carefully consulted three different curriculum guidelines, representing the work of some 20 school divisions, as we developed and refined our tests. When the SOL teacher resource guides were released by the state, we again reviewed our tests with teachers and administrators in the light of those guidelines and made what revisions were called for. In the case of the History and Social Science standards, the changes were substantial. In the other subject areas, fewer updates were needed.

Moreover, *TfHS* consults well-respected content specialists as we create and revise our tests. In the case of the reading and writing tests (English), Ms. Elizabeth Bradford and Dr. Kenneth Bradford, two well-known educators in Virginia are the test authors. Dr. S. Stuart Flanagan, a long-time and respected mathematics educator is the primary author of the *TfHS* mathematics tests. He has had extensive experience from being a test grader at the Educational Testing Service (ETS) to writing test questions for the Virginia Department of Education. (He is also Professor Emeritus at the College of William & Mary and one of the *TfHS* partners.) In the area of History and Social Science, Dr. Richard Weber, Social Studies Supervisor with the Newport News Public Schools, coordinated the primary development of the *TfHS* tests. Finally, Dr. Ron Geise, a distinguished science educator at the College of William & Mary, consulted with us on Science issues.

The content validity of the *TfHS* tests is most important for an additional reason. We find that teachers find our tests especially helpful because they show very specifically what a Standard means. Tests are an excellent medium for conveying the meaning of educational objectives or standards, as each question is a concrete instance, rather than an abstract generalization. Students find our tests valuable, for the same reason – the tests help them understand what they must learn to be proficient on the Standards. Since these two audiences do look to us for guidance, we are especially careful not to lead them in the wrong direction. We take this responsibility seriously.

Another highly relevant type of validity is ***predictive score validity***. That is, “How well do scores on our tests predict scores on the State SOL Tests?” (However, remember that the *primary* purpose of the *TfHS* tests is diagnosis, not prediction.) We have some evidence of this type from several school divisions. One small school division ran a set of regressions to predict the score on the State SOL Mathematics Test at grade 3 on the

basis of the *TfHS Mathematics Pre-Post Test* given somewhat earlier in the year. In this case, which yielded a correlation of **0.89** between the two tests, it was found that a score of 85% correct on the *TfHS* test most closely predicted a scaled score of 500 (passing) on the state test. The *TfHS Pre-Post Test* scores correlated with scaled scores on the State SOL Test **0.95** in grade 5.

In another division, administrators collected the scores shown in the in the table below. *Table 2* presents correlations between *TfHS Simulation Tests* percent-correct scores for individual students given early in spring and their later scaled scores on the State SOL Tests given in April.

<i>Grade/Course</i>	<i>Correlation</i>
Reading, Gr. 5	.61
Reading, Gr. 8	.54
Mathematics, Gr. 5	.71
Mathematics, Gr. 8	.44
Algebra 1	.76
Geometry	.73
Earth Science	.62
Biology	.74
Hist. & Soc. Sc., Gr. 8	.79
World Hist. 1	.85
World Hist. 2	.81

Table 2. Product-moment correlations between *TfHS Simulation Tests* and SOL Scaled Scores for students in two or three classes in a small Virginia school division.

A different variety of ***predictive validity indicator*** is the school-level rank correlation between the average number correct score on the *TfHS Pre-Post Tests* given in February and the percent passing the State SOL Test for that same school later in the year. We used rank correlations, as the percent-passing scores are not normally distributed. Over 31 schools, the grade-5 reading scores are correlated about **0.91** with percent passing reading. Over 32 schools the grade-5 mathematics was correlated **0.81** with the percent passing mathematics. Figures 1 and 2 are scatter-plots showing these data. The correlations are substantially higher than the correlations reported between the State SOL Tests and the *Stanford 9* tests (Reading, grade 5 was 0.78 and mathematics, grade 5 was 0.74.) Naturally, our correlations should be higher, as both the *TfHS Pre-Post Tests* and the State SOL Tests were designed to measure the same Virginia SOL Standards!

The schools in this particular sample exhibited a *very* wide range of performance. The highest scoring school in the sample had a passing percent of 93 on the State SOL Reading Test and a passing percent 88 on the State SOL Mathematics Test; whereas, the lowest school's scores were just 17 % passing the State SOL Reading Test and 12 % passing the State SOL Mathematics Test. That high scoring school scored an average of 71 % correct on the *TfHS Reading Test* and an average of 74 % correct on the *TfHS Mathematics Test* in grade 5. The lowest scoring school scored an average of 40% correct on the *TfHS Reading Test* and an average of 44% correct on the *TfHS Mathematics Test*.

Students at both of these schools fell within the range of valid measurement for the *TfHS* tests. Thus, these *TfHS* tests are suitable for measuring students' achievement over a very wide range.

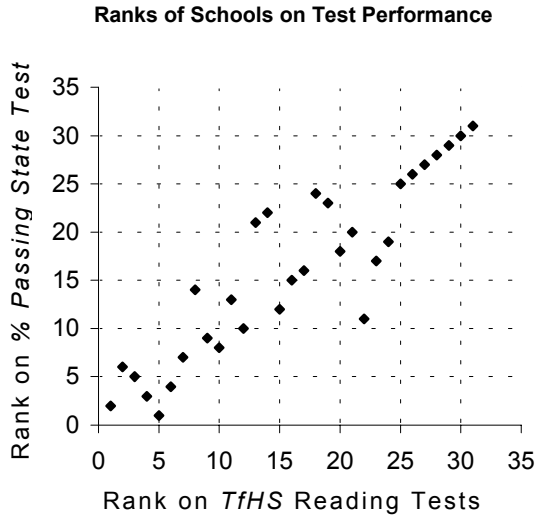


Figure 1. Relationship between percent-correct scores on the *TfHS Pre-Post Tests* in **reading** given in February and the percent passing the State SOL Tests given in the late spring of 2000. Plotted are the ranks for 31 schools. Rank 1 is the lowest.

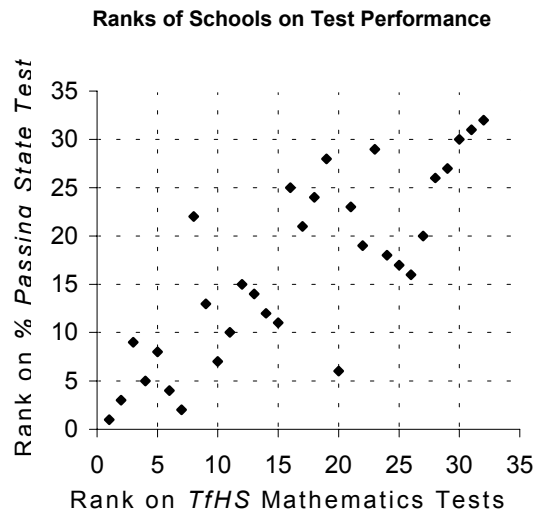


Figure 2. Relationship between percent-correct scores on the *TfHS Pre-Post Tests* in **mathematics** given in February and the percent passing the State SOL Tests given in the late spring of 2000. Plotted are the ranks for 32 schools. Rank 1 is the lowest.

Overall, we believe that *TfHS* has a set of valid and reliable assessments and would enjoy providing additional evidence to support this assertion.

David E. W. Mott
01/28/01