

January 2008

## **Proposal for a New Test**

### **Introduction**

Since the 1970 California Achievement Test (CAT) has gone out of print, many schools have been asking us about alternatives for testing. We sent out questionnaires the last two years to our customers, suggesting some possibilities and asking for ideas and feedback. One proposal was for a new test developed specifically for our plain people. There has been enough positive feedback on this idea that we are making a more detailed proposal, as follows. Please respond with your reactions, positive and negative, along with further suggestions you may have. The response we get will directly affect our decision whether or not to pursue the project.

### **The Need for a New Test**

1. The 1970 CAT is out of print, and we at Catforms have purchased Christian Light Education's remaining inventory of test booklets. We have an adequate supply of Levels 3-5 booklets available for loan to our customers. We are planning to continue to provide answer sheets and scoring services as long as people need them. We do not, however, have a supply of Level 1-2 booklets available, for testing Grades 1-3.

2. CLE is negotiating with McGraw-Hill for use of the CAT 5, a 1992 version. This test will be considerably more expensive than the 1970 CAT, and may be more restrictive as well. We will update callers with any new information we receive.

3. Other schools use the Iowa Tests of Basic Skills, or the Stanford 10. Both these tests are expensive and ordering is often impossible without the right teaching credentials. I was able to review the Stanford 10 test. It appears to be well-designed, but contains some test items most of our people would be uncomfortable using.

### **Description of Customer Base**

The proposed customer base consists of Conservative Mennonite schools ranging from a few students to a few hundred, with teachers ranging from part-time volunteers to 20- and 30-year career professionals. Also included are Amish one-room and multi room schools, other Conservative Anabaptist groups, German Baptist groups, and other interested private schools and home schools.

### **Projected Development Time**

A test of this nature would have a proposed development time of five to ten years. This would be broken down into 1-2 years groundwork, research, & planning; 2-4 years preliminary development; and 2-4 years testing and refining. Priority would be given to grades 1-3, since testing with the 1970 CAT will be unavailable for those grades first.

### **Is it possible?**

A test must have validity, that is, it must be relevant and reliable. It must also be useable and cost-effective.<sup>1</sup> I believe that such a test can be produced with the resources available to our circles. It would not be nationally normed, but it could be correlated with a national test, such as the 1970 CAT.

### **Questions Not Addressed in this Proposal**

Questions that still need to be answered address development costs and financing for such a test, and recruitment of writers, editors, and review groups. We at Catforms propose to spearhead the project, but it cannot be done without outside help and support. Again, your suggestions are welcome!

## Proposed Steps to a New Test

**Step 1. Research & education in test development** – now in progress. There has been a lot of study put into testing over the years, and there is no shortage of information about developing high quality, valid tests. The following steps will likely be modified and tuned as our research into the process continues.

**Step 2. Test blueprint** (Establish test objectives, type, format, and standards, among other things)

### A. Objectives

1. To measure students' level of mastery of concepts and basic skills appropriate to their grade level.
2. To help parents and teachers to identify both general and specific strengths and weaknesses in each individual student's knowledge.
3. To promote academic achievement and responsibility in our schools; to provide an outside standard to which schools can be accountable.
4. To help schools evaluate the effectiveness of their curriculum choices and implementation.
5. To provide a benchmark of achievement for a school from year to year.
6. To provide a test reflecting the values and academic needs of Conservative Mennonites and other plain people, unhampered by copyright and licensing issues.

### B. Type of Test

The proposed test would be a combination criterion-referenced and norm-referenced achievement test:<sup>2</sup> A specific list of concepts associated with Reading, Math, and Language (including spelling) is drawn up for each grade level. Test questions are designed to assess students' mastery of these concepts. Reporting mainly addresses the question, "How well does this student appear to have mastered this subject at this grade level?" Norm-referencing is used for setting standards for mastery, as well as for comparison with others' performance on the test. Norms would be established from the user population and correlated<sup>3</sup> with the 1970 CAT.

Mastery can be used only in a general, subject-level sense, as a single achievement test cannot have enough items to adequately establish mastery of many individual concepts. Individual concepts would be reported on a chart similar to the existing Catforms Achievement Test Analysis, showing the number of items relating to the concept, and the number of those items the student correctly answered. Teachers and others interpreting the test results would need to make their evaluations of mastery in individual concepts based on that information provided. A few questions cannot adequately test a concept, but may indicate a need to do more extensive teaching or testing on that concept.

### C. Format of the test

1. Five levels for multi grade classrooms, similar to the 1970 CAT. Grades 1-3 would write answers directly in their booklets, grades 4-12 would use answer sheets. *Note: Most current tests have a separate test level for each grade. Although this approach provides for test items that precisely target a specific grade level, it becomes very difficult to administer in multi-grade classrooms.*

2. Test items to be primarily multiple choice.

Benefits of the multiple choice format:

No subjective evaluation is required in scoring (the answer is either right or wrong, best or not best, not half-right or partly wrong).

It lends itself to detailed analysis of responses, in which even incorrect answers can provide information on the student's skills.

It lends itself well to computer scoring.

Drawback of the multiple choice format:

The multiple choice format is unable to test writing skills, including organization of thought and originality. These skills are generally beyond the scope of a standardized achievement test.

### 3. Scores provided by the Test

- Possible & raw score, from which other scores are derived
- Target raw score for each grade, to establish basic mastery of the subject (criterion-referenced)<sup>4</sup>
- Percent of that target raw score attained (PTS) (100% = basic mastery, over 100% possible) (criterion-referenced)
- Percentile scores, based on previous test results on file, possibly unpublished (norm-referenced)
- Stanine scores based on standard deviations from the norm group average, scaled from 1-9.<sup>5</sup> (norm-referenced) A stanine score of 5 is average and a benchmark for the target raw score referred to above.

This standard assumes that the average student in our customer base is achieving basic mastery in his or her grade level. This standard seems reasonable based on the fact the median percentile score of the current Catforms customer base is between 75 and 80, when tested with the 1970 CAT. What we are doing then, by basing a target raw score on the average stanine score, is building a criterion reference point on a current norm. We feel that this represents a realistic standard. It is not unattainably high, but still high enough that close to half of our current students will have to increase their performance level to reach it. This score, however, should not be used as a pass-fail criterion. A stanine score of 3 or less is considered significantly lower than the norm, and may warrant closer inspection, more testing, or remedial work.

### **D. Standards.** The proposed test would have:

- a strict focus on the academic aspects of Reading, Math, and Language.
- no references to elements most of our customers would find objectionable, such as humanism, evolution, politics, television, radio, musical instruments, movies, concerts, organized sports, amusement parks, racing, alcohol, tobacco, etc.
- high quality line art for illustrations; no artwork depicting people or of a frivolous or comic nature.
- no frivolous, so-called politically-correct, or historically revisionist reading passages.
- passages designed to test reading comprehension that are original or else so little-known that no student has an advantage in answering test items because of familiarity with the passage.
- items that test a single concept as much as possible, in order to accurately assess mastery of that concept. An exception: Math story problems that require a student to read, assess, isolate steps to a solution, and accurately process those steps.
- no trick questions, and no attempts at humor. All items are to be as straightforward as possible, so that the student spends his time working on solutions, not figuring out the questions.
- test items written and evaluated, as much as possible, according to standard practice as defined by the Standards for Educational and Psychological Testing.<sup>6</sup>
- a transparent and well-documented development process.

### **Step 3. Detailed learning objectives, arranged in detailed scope & sequence**

Many of the same concepts would be covered as tested in 1970 CAT. Concepts would be tested in grade levels suggested by the scope and sequence of CLE, Rod & Staff, and other Mennonite curricula. Comparisons would be established with the Reading, Math, and Language sections of the Core Knowledge Sequence<sup>7</sup> as a nationally-respected guideline for private schools and home schools. A detailed Scope & Sequence table would be prepared, showing concepts as tested compared to their introduction in various curricula and the Core Knowledge Sequence.

#### **Step 4. Item construction and banks**

The work of developing test items has become significantly easier in the past several years with the advent of test writing software. Multiple versions of the same item can be created simultaneously, with wording changed, different numbers used, and answers shuffled. With such tools, we would like to see a general test similar in length to the 1970 CAT and possibly also specific tests with enough items in a particular subject, say 6th grade math, for further testing and evaluation if desired.

Items are to be well-documented and described. Classifications for items include subject, general concept area, specific learning objective, keywords, grade level, thinking process required according to Bloom's taxonomy<sup>8</sup>, and difficulty level relevant to the grade level. Some of this information will be primarily used in the selection process. Items will then be evaluated, saved in item data banks, and used as needed for the general test or the individual tests suggested above.

Many test publishers have ready-made item data banks available, where individual test items may be utilized in a user-generated test. More research is needed about the availability and usability of such data banks. Usage rights are probably the biggest question, followed by cost. If use of a commercial item bank turns out to be practical, we may save much time, obtain quality test items, and retain our own selection process.

#### **Step 5. Item testing & analysis**

The development of high quality items is essential to a valid test. There are many rules governing the writing of multiple choice items in particular, although we don't have room to elaborate here.<sup>9</sup>

Items must be written in accordance with these rules. These items must then be tested carefully on small populations of students. A good test item must be relevant and appropriate to the grade level, discriminate correctly between high-scoring and low-scoring students, and have plausible distractors. All these criteria are impossible to analyze accurately without actual testing.

Even good item writers need to produce about twice as many items as will be used in the completed exam.<sup>10</sup> The elimination and rewriting process is critical to success.

#### **Step 6. Pilot testing & correlation**

After the test items have been written and individually analyzed, sections of the test must be compiled and evaluated again, by testing them on larger groups of students. These subtests must be internally correlated and evaluated for reliability. As with the writing of individual items, there are also specific procedures to follow for validating subtests.<sup>11</sup>

#### **Step 7. Review & revision**

After the correlation results are calculated, the sub-tests must again be examined. Are there individual items that skew the test by not accurately reflecting overall results? After items have been weeded out, are there concepts that are no longer adequately represented? Final page layouts and illustrations not directly part of individual test items must be designed and reviewed for quality and clarity.

#### **Step 8. Large-scale testing & norming**

After as much tweaking and tuning has been done as possible, a "beta" version of the test would be released. This would target as many schools as possible and these test results would be subject to the most scrutiny of all. Simultaneous administration and correlation with the 1970 CAT would take place. Establishment of norms and target test scores would also be done with the results of these tests. In all testing procedures, students should, if possible, not be informed of the experimental nature of the test, to avoid possible performance changes compared to taking a "real" test.

#### **Step 9. Publication**

The final step in initial test development is publication and distribution. After that, work on revisions, corrections, and new editions will need be part of the ongoing program.

## End Notes

1. James Cangelosi, Designing Tests for Evaluating Student Achievement (White Plains, NY: Longman, 1990) pp. 27, 36
2. For one description of criterion referencing vs. norm referencing, see Norman E. Gronlund, How to Make Achievement Tests and Assessments, 5<sup>th</sup> Ed, (Needham Heights, MA: Simon & Schuster, 1993) pp. 12-15
3. The term “correlation” is used multiple times in this document. For a discussion of this subject, see Measurement and Evaluation in Psychology and Education, 6<sup>th</sup> Ed, by Robert M. Thorndyke (Upper Saddle River, NJ: Prentice-Hall, Inc, 1997) pp. 43-48
4. The concept of a target test score is somewhat unusual in achievement tests. Please share your thoughts about this idea in particular.
5. Cangelosi, 221
6. American Educational Research Association (AERA), et al, Standards for Educational and Psychological Testing (Washington, DC: AERA, 1999)
7. Core Knowledge Foundation, Core Knowledge Sequence: Content Guidelines for Grades K-8 (Charlottesville, VA: Core Knowledge Foundation, 1999)
8. From Benjamin S. Bloom, ed., and others, Taxonomy of Educational Objectives: Cognitive Domain (New York: David McKay Co., Inc, 1956), pp. 201-207. Paraphrased slightly, the levels of thinking and understanding according to this classification are: *knowledge*, or simple recall of facts; *comprehension* of the meaning; *application* of the facts in specific situations; *analysis*, or breaking down the material into its parts; *synthesis*, putting parts together into a whole; and *evaluation*, judging the value of a thing for a specific purpose.
9. Whole books have been written on the subject. One in particular is Developing and Validating Multiple-Choice Test Items, by Thomas M. Haladyna (Mahwah, NJ: Lawrence Erlbaum Associates, Inc., 2004). Also see Thorndyke, pp. 444-473, for both general and specific guidelines on multiple choice and other item formats.
10. Thorndyke, 476 (I know I read it more specifically somewhere, but cannot now lay eyes on it)
11. Thorndyke, 476-483

## Other Resources

- Downing, Steven, and Thomas Haladyna, Ed, Handbook of Test Development (Mahwah, NJ: Lawrence Erlbaum Associates, Inc., 2006)
- Brown, Frederick, Principles of Educational and Psychological Testing, 3<sup>rd</sup> Ed (New York: Holt, Rinehart and Winston, 1983)
- Airasian, Peter, Classroom Assessment, 3<sup>rd</sup> Ed (New York: McGraw-Hill, 1997)