KeyMath–Revised: A Diagnostic Inventory of Essential Mathematics (Connolly, 1998) is an individually administered, norm-referenced test. The 1998 edition is a normative update only. The basic testing materials consist of two easels that contain testing items and directions for presenting and scoring items. Four uses are suggested for the test: (1) instructional planning, (2) comparison of students, (3) evaluation of educational progress, and (4) curriculum evaluation.

There are two forms of KeyMath–R (forms A and B), and each contains 258 items. Total math performance is divided into three areas. The area of basic concepts comprises three subtests: numeration, rational numbers, and geometry. The operations area consists of addition, subtraction, multiplication, division, and mental computation. The area of applications contains items assessing measurement, time and money, estimation, interpretation of data, and problem solving. Each subtest in turn, comprises various domains. A domain is a subdivision of a subtest; for example, in the measurement subtest, there are four domains: comparing, using nonstandard units, using standard units of length and area, and using standard units of weight and capacity. The subtest of rational numbers consists of three domains: fractions, decimals, and percentages. There are 3 or 4 domains per subtest (for the 13 subtests), for a total of 43 domains. Written computation is permitted only on some of the subtests in the operations area.

Scores
For each subtest, both percentiles and standard scores (with a mean of 10 and a standard deviation of 3) are available. For area performance and total test performance, six derived scores are provided: standard scores (with a mean of 100 and standard deviation of 15), normal-curve equivalents, stanines, percentiles, age equivalents, and grade equivalents. Finally, KeyMath–R provides a rather unusual score for domains: Domain-performance scores divide student performances into strong (top quartile), average (middle quartiles), or weak (bottom quartile). The usefulness of domain scores for instructional planning is unclear. A computer program is available to convert raw scores and to construct student profiles.

Norms

The 1998 normative update was completed in conjunction with normative updating of the Peabody Individual Achievement Test–Revised, the Kaufman Test of Educational Achievement–Revised, and the Woodcock Reading Mastery Tests–Revised. The sample for the normative updates was 3,184 students in kindergarten through grade 12. A stratified multistage sampling procedure was used to ensure selection of a nationally representative group at each grade level. Students in the norm group did not each take all five tests. Rather, one fifth of the students took each test, along with portions of each of the other tests. Thus, the norm group for KeyMath consists of about 600 students. There are as few as 91 students at three-year age ranges. Because multiple measures were given to each student, the authors could use linking and equating to increase the size of the norm sample.
Reliability

Alternate-forms reliability was estimated by retesting about 70 percent of the students in grades K, 2, 4, 6, and 8 at two- and four-week intervals. However, Connolly does not report estimated reliability by grade; rather, he reports pooled (across-grade) coefficients. Only the total score may be sufficiently reliable for making important educational decisions for students; all subtest and area estimates of reliability are less than .85.

Split-half reliabilities (using odd–even splits and Spearman–Brown correction) were also estimated by grade and age. For students in kindergarten through second grade, total scores are consistently reliable enough to use in making decisions for individuals; area subtests fluctuate, so the test user must determine whether a particular age–area combination is sufficiently reliable for interpretation. After second grade, area scores have acceptable reliability, and total scores have excellent reliability; however, because basal and ceiling rules were applied to the test scores, the obtained split-half estimates are likely to be inflated.

Another method of estimating reliability based on item-response theory was used. The results of this analysis are essentially the same as those obtained using split-half estimates.

1 1. We believe that the reliability estimates based on age are somewhat misleading because several age groups are contained within any grade. Consequently, the range of ability would probably be extended.
No reliability data are provided for domain scores. No stability coefficients are reported, although stability can be inferred from the alternate-forms reliability coefficients.

Validity

Little evidence of construct validity is presented. What is offered is a demonstration of mean-score progressions from grade to grade. No evidence of concurrent validity is presented. However, for most achievement tests, these indexes of validity are less important than evidence of content validity. Limited evidence for KeyMath–R's content validity comes from the careful development of a table of specifications to guide item development. As is always the case, however, test users should inspect the test's content to make sure that it conforms to the curriculum followed by the students who are being assessed.

Summary

Based on its full title and the claims made in its manual, KeyMath–R is intended as a diagnostic test. The standardization of the test appears to be generally adequate. For grades before third grade, only the total score is sufficiently reliable for diagnostic purposes; for third grade and later, area and total scores are sufficiently reliable. Because of the low reliabilities of subtests and the absence of reliability information for domains, users should avoid making inferences about a student's instructional strengths and
weaknesses based on subtest and domain scores. Of the four uses that are proffered for
the test, no evidence of KeyMath–R's validity for instructional planning, evaluation of
educational progress, or curriculum evaluation is presented. Some evidence for the
validity of comparing students' global performances is presented.