Validity of the Grade 4 and Grade 8
English Language Arts and Mathematics Assessment

- The demonstration of validity of the statewide assessments in grades 4 and 8 relies heavily on their agreement with the State Learning Standards. In each step of the test development process, expert judgment is used to complement empirical analyses to ensure that each item is related to the appropriate standard. Once items are assembled and field tested, the standard setting study builds the test's capacity to distinguish students who have achieved the Learning Standards from those who have not. The study involves an empirical analysis of items that are designed to measure each of the Learning Standards as well as deliberative expert judgment.

- Extensive analyses of the correlations of test questions to the test scores, and of the fit of the test questions to the test as a whole demonstrate that each question measures the desired Learning Standard.

- To demonstrate the appropriateness of the Grade 4 ELA test for all populations, analyses were made of whether the test measures the same trait for students with disabilities and for English Language Learners as for all other students. Analyses of the three June tests will be available soon.

- Predictive validity cannot be assessed without a longitudinal data base. Predictive validity traditionally rests on correlation. Passing rates could differ dramatically between instruments that correlate well, if the instruments measure the same trait. Nevertheless, large field test analyses and the limited small sample department reviews suggest comparable passing rates between the eighth grade instruments and the single June 1999 Regents examinations (58 percent passing rate on English and 35 percent passing rate on Mathematics A). Confidence in the relationship among tests for students of different developmental levels relies on the relationship among the Learning Standards designed for those levels and the empirical evidence that each question measures the Learning Standard of interest.

- Reliability concerns the agreement of the observed score to the true score, or the replicability of the student's performance across measures and time. Naturally, the greater the breadth of the measure, the more test questions, the greater the reliability. Reliability is best achieved by evaluating performance based primarily on the whole test before considering smaller portions of the test. (See Decision Models attachment.) The internal reliability coefficients for the four examinations are quite respectable: .88 (ELA-4), .90 (ELA-8), .93 (Math-4), and .94 (Math-8), especially for tests using items of various formats.

- The empirical analysis of bias made of each test question assures that for large population subgroups the questions each measure the same trait. The analyses are used to screen items from test forms before they are administered.

- Pattern scoring determines the scale score that allows for the greatest probability of observing the particular configuration of correct and incorrect answers each student demonstrates. A description is attached.