An Evaluation of Speededness and Fatigue Factors on the Grade 4 English Language Arts Examination, February 2000

Gerald E. DeMauro
April 26, 2000

An important issue in test performance is the extent to which observed performance is influenced by factors unrelated to the construct being measured. In particular, for examinations of younger children, attention is given to the issues of speededness (was there enough time to answer the questions) and to fatigue factors. Field test analyses of these factors suggest that neither one of them is a systematic threat to score validity; that is, that neither of these factors is related to or detracts from the inferences one can make from the children's test scores.

Nevertheless, an evaluation was made of the data available from the operational examination of the Grade 4 English Language Arts test (ELA-4) administered in February 2000. This examination is administered over a three-day period. It consists of 28 multiple-choice questions administered on the first day, and four cluster or open-ended questions. The first two are administered on the second day, and the fourth is administered on the last day. The third cluster contains three parts, one of which is administered on the third day and the other two are administered at the end of the second day. We hypothesize that, although these questions may differ in difficulty (indeed, the more difficult questions are often better-placed at the end of a timed test section), that the increased difficulty is not related to non-model or non-construct influences such as fatigue or speededness.

To test this hypothesis, the initial data tape was evaluated from the test administration. The test questions were divided according to the stimulus with which they were associated. Essentially, this divided the questions into four sets of multiple choice questions, of seven, five, six, and five questions each. These had mean values of 6.12, 3.62, 4.18, and 3.88.

The four cluster totals were divided into maximum possible points of four, three, three, and four points each. The observed means were 2.31, 2.06, 2.06, and 2.13, respectively.

To determine whether there were systematic influences on scoring that were unrelated to the measure of interest, the observed points achieved were divided by the number of possible points for each of the four multiple-choice and four cluster sections. This controls for observing higher influence of any

section simply because the score contributes more points to the total. All questions were then totalled, and the contribution of each of the eight sections to that total was estimated by correlations.

The four multiple-choice correlations, all from the first session, were: .659, .610, .577, and .651, respectively. The last section had the second highest relationship to the overall total, surpassed only by the first section, indicating no systematic source of bias or corruption related to fatigue or speededness.

The four cluster sections exhibited correlations of .765, .803, .841, and .810, respectively. The last section again had the second highest relationship to the overall total, again suggesting no systematic source of bias in relation to speededness or fatigue on the last day. Moreover, the highest correlation of all cluster scores was observed for cluster 3, which contained the last items administered in session 2. Again, there is no suggestion of speededness or of fatigue.

When research tapes are finalized, these analyses will be repeated. However, there is no indication of speededness, fatigue, or systematic corruption of the measurement properties of the test in relation to the placement of the test sections.