Over the past two decades, the use of standardized tests in the nation’s schools has increased sharply. According to a report prepared for Congress more than 10 years ago by the Office of Technology Assessment, “The rise in testing reflects a heightened demand from legislators at all levels—and their constituents—for evidence that education dollars are spent effectively. Holding schools and teachers accountable has increasingly become synonymous with increased standardized testing” (U.S. Congress, Office of Technology Assessment [OTA], 1992, pp. 3–4). This trend culminated with the passage of the No Child Left Behind Act of 2001 (NCLB) in 2002; that law requires that each state administer census assessments of reading and mathematics proficiency in grades 3–8, and at least once during grades 10–12, beginning with the 2005–2006 academic year. Additional assessments of science are required by academic year 2007–2008.

This increase in emphasis on the use of standardized tests to assess achievement in the schools has led to a corresponding increase in interest in the results obtained by national and international surveys of educational achievement that provide aggregate results for the nation and internationally. Chief among those are the National Assessment of Educational Progress (NAEP) and the Third International Mathematics and Science Study (TIMSS).

The NAEP is a widely respected indicator of educational performance (Beaton & Zwick, 1992), with a scale that offers national comparability and information on change over time.

---

1The opinions expressed in this chapter are those of the author and not necessarily of the University of North Carolina at Chapel Hill.
NAEP has proven to be a valuable tool to track and understand educational progress in the United States. It was created in 1969 and is the only regularly conducted national survey of educational achievement at the elementary, middle, and high school levels. It was designed to be an educational indicator, a barometer of the Nation’s elementary and secondary educational condition. NAEP reports group data only, not individual scores. (U.S. Congress, Office of Technology Assessment, 1992, pp. 30–31)


The administration of NAEP and the presentation of its results are unusual in that no scores are assigned to individual examinees. Instead, a complex sampling design is coupled with the models and methods of item response theory (IRT) to yield estimates of statistics that describe the population distribution of proficiency: the mean, various quantiles, and the percentages with proficiency in regions on the scale known as achievement levels.

TIMSS is the current incarnation of a series of international comparative studies conducted by the International Association for the Evaluation of Educational Achievement (IEA) since its inception in 1959 (Mullis et al., 1997). Whereas TIMSS releases scores on several scales, its primary reporting, like that of NAEP, is for large demographic groups—in the case of TIMSS, nations.

For a number of reasons and purposes to be discussed in subsequent sections, there have been several attempts to link the results and scales of NAEP and TIMSS (and other international assessments) to each other (Beaton & Gonzalez, 1993; Johnson, 1998; Johnson & Siegendorf, 1998; Pashley & Phillips, 1993) and to statewide assessments (Ercikan, 1997; Linn & Kiplinger, 1994; McLaughlin, 1998a, 1998b; Waltman, 1997; Williams, Rosa, McLeod, Thissen, & Sanford, 1998). These linkages differ from many other more common applications of equating and from the construction of concordance tables, in that for NAEP (and TIMSS, for the most part) there are no individual scores to put in a concordance or “cross-walk” table. Instead, the goal is to use the results from the administration of some other assessment to make estimates of the aggregate results that an assessment like NAEP might produce. The idea that statistical linking might be used to accomplish that goal is relatively new, having its genesis in systematizations of test linking by Mislevy (1992) and Linn (1993); those descriptions of alternate forms of test linking are described in Section 16.1.