Chapter 5.3
Introducing a Computer-Adaptive Testing System to a Small School District

Timothy Pelton
University of Victoria, Canada

Leslee Francis Pelton
University of Victoria, Canada

ABSTRACT

A computer-adaptive test (CAT) is a relatively new type of technology in which a computer program "intelligently" selects and presents questions to examinees according to an evolving estimate of achievement and a prescribed test plan. A well written CAT can be expected to efficiently produce student achievement estimates that are more accurate and more meaningful than a typical teacher-generated paper and pencil (P&P) test with a similar number of questions. Although this method of testing sounds good in theory, many schools and districts are waiting for positive examples of practical applications and observable benefits before adopting a CAT. This chapter begins by describing the essential elements of meaningful measurement in education and the features of a typical CAT. Next, we describe the Measures of Academic Progress (MAP) system of the Northwest Evaluation Association (NWEA; 2004) and observations made during the introduction of this system into a small semirural school district. Finally, as independent observers, we provide a set of recommendations to help guide other districts as they consider the potentials of implementing a CAT system to guide instruction within their schools.

BACKGROUND

A Need for Meaningful Measurement

Underlying the implementation of a CAT is the assumption that there is a meaningful interval scale of development and growth upon which the curricular strands or constructs identified in the test specifications can be mapped.

To support meaningful measurement, construct-linked items must fit within a stable hierarchy of difficulty, such that if a person were to be presented with the entire set of items for a particular construct starting from least difficult and progressing to most difficult, the resulting response vector would consist of a sequence of almost entirely correct responses followed by a sequence of almost entirely incorrect responses (i.e., approximating a Guttman response vector; Guttman, 1944). The expectation of an approximate hierarchy of item difficulties is not a great barrier—master teachers implicitly use this type of internal hierarchy as they adjust their diagnostic questioning to more accurately understand the achievement level and needs of individual students—but the requirement is important and needs to be attended to when examining a test item bank. Without this hierarchy of construct-linked items, a valid adaptive test is not possible.

Given a well refined hierarchical set of construct-linked items and a substantial set of responses from members within the target audience, it is possible to obtain a meaningful measurement scale by analysis using item response theory (IRT) or the Rasch measurement model (Lord, 1980; Pelton, 2003; Rasch, 1993; Wright, 1991). Difficulties inevitably arise when an attempt is made to build a unitary assessment device for an entire curriculum that in reality is a very complex collection or network of constructs. Here, pragmatism and optimism tend to lead test developers to implicitly assert that each of the curricular strands approximates a well defined construct and that these constructs load primarily onto a common scale in such a way as to maintain a meaningful hierarchy of difficulty both within the constructs and between them.

The limitations of a finite calibrating data set and the pragmatic binding together of a complex domain of constructs onto a single achievement scale mean that the assumptions of the measurement model being used (i.e., Rasch or IRT) are being broken, and therefore the underlying domain scale should really be described only as quasi-interval. Although this type of measurement scale is not perfect—indeed, the error levels are likely to be understated (Pelton, 2003; Pelton & Bunderson, 2003; Pelton & Francis-Pelton, 2004)—it can be argued that the units have a relatively stable meaning and, as such, allow for more appropriate mathematical and statistical analysis than traditional test score values.

What is a CAT?

CAT software typically uses an evolving estimate of achievement level (starting from a random guess or based upon a teacher-generated estimate) to locate items in an item bank that are likely to be effective in testing the student's knowledge and skills (i.e., at or near the student's achievement level). The student response (correct or incorrect) is then incorporated into the information used to generate the next achievement estimate. This iterative process typically continues until the error estimate associated with the current achievement estimate diminishes to an acceptable level (i.e., increasing the reliability of the test until it reaches an acceptable level), or until the test specification requirements have been met.

The test specifications are used in CAT systems to support validity by ensuring that a sufficient number of questions are asked for each of the specified constructs (curricular strands) in the domain and that the test includes sufficient coverage of the cognitive skill levels. With some test specifications, longer CATs may also be able to generate valid estimates of achievement for each