Institute of Education Sciences
National Center for Education Statistics

NCES/NISS Task Force on Computer Adaptive Testing for Longitudinal Studies

January 2008

Executive Summary

The Task Force was convened to consider the utilization of Computer Adaptive Testing in NCES longitudinal studies in general and in HSLS-09 in particular. Computer Adaptive Testing (CAT) is distinguished by the adaptive selection of items for an individual test-taker based on previously collected external information and on responses to earlier items.

> *The challenge in any educational assessment is simultaneously to maximize the information gathered through testing both for individuals and for groups, to optimize the efficiency of the testing process, and to create a resource database for the study of educational issues, practices and outcomes.*

The Task Force considered the consequences of individualized selection of items based on a test-taker's response to earlier questions. Aspects of assessment that are affected include the basic structure of items and item sequences, the process of item development, scoring and score analysis. Aspects of implementation that are affected involve the mode of test presentation, data capture, data storage, data file structure and curating of data. Therefore, the Task Force examined each of these aspects of CAT as they would be important for longitudinal studies such as NCES undertakes. Following extensive deliberations, the Task Force reached four primary conclusions.

Conclusions

The overall conclusions of the Task Force are that:

- CAT is *feasible*, because of technological advances, as well as students' facility with computers;
- CAT is *desirable*, reflecting the evolution of education mechanisms, practices and goals;
- CAT is *effective*, often uniquely so, in the face of time limitations and other practical constraints;
- HSLS-09 is an *appropriate entry point* to CAT for NCES, because of its longitudinal nature and because its assessments are not used for evaluative purposes.

Principal Findings

Adaptive Computer-based Testing

CAT is a particular form of computer-based testing (CBT), which is now a mature, well-tested set of technologies. In particular, "paper" testing does not offer advantages in regard to issues such as:

1. Use of interactive items,
2. Opportunity to assess e-learning and e–learning processes,
3. Options to present information over time to evaluate the new information and reevaluate earlier information,

4. Opportunity to assess attentiveness and /or the incorporation of mechanisms to improve attentiveness, or
5. Security of information, both to minimize data loss and to immediately recapture lost or inconsistent information (rather than by means of recalls or repeated visits), with careful attention to computer security especially *vis-à-vis* information transfer.

CAT is appropriate for broad assessments requiring comparable precision over a fairly wide range of performance, or being widely used for analysis, including analyses of subsets of data or of subgoups of the population. Longitudinal studies typically have both these attributes. Another principal advantage of CAT is efficiency in terms of students' limited assessment time (or an equivalent improvement in scoring precision). Various types of adaptive designs (i.e., two-stage adaptive designs, item adaptive designs, testlet adaptive designs, or variants of these) carry different relative advantages. The choice in a particular instance depends on the specific assessment goals, the structure of the test and the time available for the test administration.

A major advantage of CAT is that adaptation can occur with respect to both previous responses as well as information from external sources (student background and history, previous assessment results,…). By integrating this external information into the design, CAT may be able to provide acceptable precision even in the ever-shorter available assessment times, *e.g*., by sharpening the "zero-th" stage process to identify approximately correctly the student's performance level to initiate testlets closer to the actual (true) performance level.

Computerized adaptive testing requires an early commitment to an adaptive form. Item construction and item calibration in computer-administered form need to precede or proceed in parallel with formulation of the adaptive structure and then the algorithm development. Software development also requires early decision about platform(s), security requirements, ancillary features (such as monitoring attentiveness), data transmission verifications and the ultimate database structure. The ultimate benefit with this early preparation is streamlining the compilation of the data base by minimizing difficulties in data transmission, data editing and data base creation.

CBT, including CAT, can be conducted on-site, or remotely using the World Wide Web. The use of secure internet data *transmission* is usually crucial and feasible. In contrast, web-based test *administration* (remote or on-site) is seriously flawed, carrying with it the potential of many costly difficulties that have not yet been resolved. These range from coordination of administration to security of items and responses, authentication of test-takers' identities, adequacy/equivalence of multiple platforms, as well as disruptions to internet performance and other problems outside the control of the test administration staff.

Adaptive Testing for STEM Constructs and Facets

Some aspects of educational assessment are largely independent of the mechanism of administration, be they CAT or other mode. These include determining which constructs and facets to measure. In general, choice of aspects of constructs to measure can be justified in terms of:

1) providing the largest amount of *additional* information, given information from other sources,
2) reflecting specific covariates of general importance and covariates that differ among subpopulations of particular interest.
3) directly relating to primary inferences to be drawn both for cross-sectional and for longitudinal analysis objectives, and
4) having educational relevance, both in the source for the construct and in the potential for actions to be taken in consequence of the assessment results.

Some constructs seem to be uniquely or distinctly more easily measurable via CBT or CAT, such as mastery of processes involved in e-learning, of complex reasoning processes involving the evaluation of the value and validity of individual pieces of information, or of complex tasks for which the subtask sequence depends on the degree of complexity the student is able to recognize.

Any assessment in STEM settings is complicated by lack of a universal science construct. Science is not monolithic, nor are reasoning skills and interest levels fully shared across scientific areas. It has yet to be demonstrated that context-free assessment of scientific skills, especially the capability to reason with scientific information, can be accomplished. Breadth across sciences and depth in a particular area of science are both important.

Elucidating the interplay of any of these with a host of covariates (family structure, socio-economic status, influence by family members, friends, teachers, the popular media,…) seems to demand CAT, especially given the longitudinal nature of many of the research and policy questions.

Using CAT, a hybrid design in the STEM setting that addresses the issues just discussed is possible and feasible. One solution is to use a multi-science single context, for example an environmental problem, for several testlets. The first of these can be more general and presented to all students; subsequent testlets can be individualized to be specifically in each student's designated (preferred) domain.

Mathematics is essentially linear through the "college-track" algebra-to-calculus level[1] for the US "college-track" algebra-geometry-trig-calculus sequence, although this is not the case for other tracks or for other (e.g., foreign) curricula. As a result, mathematics poses fewer difficulties than science in a high-school setting.

Use of CAT for STEM raises serious but manageable issues of item sources and assessment design. Concerning the former, existing item pools, (e.g., PISA, NAEP, NELS and others) are extensive and well calibrated. With careful selection of items, these can serve as a valuable, but not sole, source of suitable items, noting that items with multiple plausible distractors are required for partial scoring. There is, however, an important caveat: *Item calibration for a traditional paper-pencil multiple choice test is not automatically sufficient to calibrate the same items for inclusion in computer adaptive tests, although it is often possible to get close enough to allow equating.* For constructs that can be assessed uniquely by computer-based or computer-adaptive testing, new items need to be constructed, with emphasis on responses that can utilize a scoring model that gives partial credit.

In terms of assessment design, evidence-centered design (ECD) creates a structural approach to the design, implementation and delivery of assessments. These principles form the basis for a "best practices" approach that is fully applicable to longitudinal studies and that can directly incorporate computer adaptive test structures.

Implications for HSLS-09

HSLS-09 meets all the principal criteria for use of computer adaptive testing. For students, HSLS-09 is a "low stakes" test: there is little incentive to cheat, but correspondingly non-negligible likelihood of inattentiveness. The test administration time for HSLS-09 is severely limited. The two planned assessments (early 9th and late 11th grades) can be done using a single platform and supporting software. Extensive auxiliary information is to be incorporated into the database, which leverages the strengths of CAT by allowing pre-categorization of both each student's specific scientific strength (preferred domain) and of the most appropriate level for entry into a CAT.

Link to the Full Report

[1] AP statistics may be the only numerically significant departure from the algebra-geometry-trigonometry-calculus path.