Institute of Education Sciences
National Center for Education Statistics

NATIONAL INSTITUTE OF STATISTICAL SCIENCES
EXPERT PANEL REPORT

COMPUTER ADAPTIVE TESTING

TABLE OF CONTENTS

NATIONAL INSTITUTE OF STATISTICAL SCIENCES

TASK FORCE REPORT ON COMPUTER ADAPTIVE TESTING

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 <u>not automatically</u> 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.

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

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.

NATIONAL INSTITUTE OF STATISTICAL SCIENCES EXPERT PANEL REPORT

PREFACE

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 met twice in person to develop a document that presented the opportunities for CAT within the context of NCES assessments and longitudinal studies and to provide examples to illustrate how items might be implemented in these contexts. The Task Force also considered the merits of moving to a computer-based assessment/survey process and made recommendations to NCES about adopting such a process.

COMPUTER–ADAPTIVE TESTING FOR LONGITUDINAL STUDIES

OVERVIEW

CHARGE TO TASK FORCE

The Task Force was convened by NISS 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 aspects of CAT that would be important for longitudinal studies such as NCES undertakes. After extensive deliberations the Task Force reached the following conclusions. The details contributing to these conclusions are detailed in the body of this report.

CONCLUSIONS

- 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.
- HSLS-09 is an *excellent opportunity* to take advantage of the strengths and flexibility of CAT to integrate administrative records, prior test performance results and interest information into the adaptive testing algorithm.

SECTION I

§ 1.1 COMPUTER-BASED AND COMPUTER-ADAPTIVE TESTING

Computer-based testing (CBT) is now a well-established technology. In the beginning computer-based data acquisition was considered primarily in terms of translating paper/pencil interview and test instruments to the keyboard or touch screen. At the present time, CBT methodology has broadened, for example, to update and compare existing data bases virtually simultaneously and to extend data recording beyond item responses to timing, vacillation between responses, distraction from the test at hand, etc. Available hardware has expanded to allow data acquisition via computer touch screen, hand-

held device, audio query and voice-capture/analysis, with options of test administration remotely or on-site in either continuous or segmented mode. The panoply of alternatives continues to widen, promising for the future new opportunities to alter the testing process from the schedule to the test design to the constructs selected for measurement to the mechanics of delivery.

ADAPTIVE TEST STRUCTURES

Adaptive testing attempts to come closer to an ideal of assessment by an individual examiner who poses each question based on all the information gathered in the examination thus far to reach a refined assessment of the examinee. Adaptive test structures are characterized by an initial stage (one or more questions) followed by one or more decision points where the next question or group of questions is selected based on earlier responses. The strength of an adaptive structure is to "customize" the test, rapidly coming to focus on items close to the examinee's actual proficiency level and eliminating time otherwise spent on items that are patently too difficult or too easy. The potential risk with an adaptive structure is of "misrouting" an examinee due to early responses that were either accidentally or atypically wrong or serendipitously correct. The obvious trade-off is between an extensive initial stage to lower the probability of misrouting and the fraction of highly informative items for that examinee from later stage(s) and the precision of the final score. Some adaptive structures allow for correction of misrouting when it does occur.

As practiced, adaptive testing uses responses to items to determine the item selection for subsequent items. For paper-and-pencil tests, this form of testing is usually implemented via a two-part test with a single version of the first part, but with several different versions at different levels of difficulty for the second part. For the second part of the test, the version given to a student is determined by the (rapidly scored) result from the first part.

Adaptive testing strategies utilized in other implementations have primarily had either of two general structures: two-stage split test (like paper-and-pencil adaptive tests) or item-by-item adaptation. Alternative structures include subdividing the test into segments or testlets, each of which can be designed as a mini-adaptive test with one or more items at each stage. The two-stage split test requires many items in the first stage to avoid an unacceptable frequency of misclassification for the second stage; and correction for the occasional misrouting is not possible. The item-by-item adaptive test requires a large item pool and leads to an extraordinarily (often prohibitively) large number of possible paths, and commensurate expense in calibration time and cost.

Testlet adaptive structures subdivide the whole test into segments or testlets with adaptation occurring within each testlet. For example, a 20-item test could be divided into five testlets, possibly with distinct content or specific context for each, although this is not necessary. Then each testlet could follow a two-stage adaptive design with a pair of questions at the first stage to determine the level of difficulty for a second pair at the second stage. Alternatively, the 20-item test could be split into ten pairs with the selection of the second question dependent on the response to the first. However, in any case, each subsequent testlet presents the opportunity to correct any misrouting that occurred previously and also to allow for different proficiency levels for the distinct content areas covered in the test. A testlet structure can also allow adaptive testing with adaptation both within and between context/concept testlets. The adaptation within testlet follows the conventional paradigm to utilize the second stage to refine the testlet score within the narrower range, determined by the first stage testlet response(s).

The adaptation between testlets allows (modest) modification of the second-stage entry point with each testlet so that an aberrant response on a single testlet does not control the entire assessment. (See **§ 3.3**, *2.5 Stage Designs* and *Testlet Diagram and Scoring*.)

COMPUTER-BASED ADAPTIVE TESTS

Computer adaptive testing (CAT) exploits CBT technology by assembling the advantages of adaptive testing and embedding these in this adaptable and flexible context. The item selection and sequencing are accomplished by algorithm, and allowance can be made for patterns of responses as well as individual responses - just as the idealized individual examiner might do. CAT is far more flexible than paper-and-pencil adaptive tests: adaptation can occur one or more times during the test, and adaptation algorithms can be more complex, for example, incorporating time required to reach a correct solution as well as the final response and item sequencing as well as item selection can be dynamic.

While the advantages of employing CBT are chiefly the greater range of item formats and the improvement in data capture and data handling, the additional advantages of CAT are the greater flexibility and real time adaptation can result in greater efficiency without introducing any new difficulties.

§ 1A HSLS-09 OBJECTIVES AND CONSTRAINTS

HSLS-09, like predecessor longitudinal studies, will track a cohort of students entering high school (grade 9) and follow them through their high school course-taking trajectories on into their post-secondary activities and/or academic programs.

HSLS-09 is intended to provide a database on the decision-making process of high school students in regard to post-secondary education, with a particular focus on persistence and attainment in STEM majors and careers. Surveys will begin at high school entry and extend into the post-secondary years through a total of at least four surveys, two during high school and two subsequently; in-school surveys will incorporate math and science testing. The principal purposes of HSLS-09 are to observe dynamic processes; there is a special interest in understanding the correlates of the choices that students make, particularly those related to persistence and attainment in STEM (post-secondary) majors and careers.

The constraints for HSLS-09 are especially severe with one-hour total contact time with each student at each assessment. This therefore demands rapid focus of the testing at the individual student's level and extracting maximal information from a small number of items. The testing schedule provides for:

- 2 assessments: early 9th grade assessment (fall 2009), late 11th grade (spring 2012)
- 5 administrative record collections: prior to each assessment (2009 & 2012), at the conclusion of high school (spring 2013), also 2 and 6 years post class graduation (2015 & 2019).

Each assessment will be limited to one hour, of which 40 minutes in total may be devoted to mathematics and science, and 20 minutes allocated to an attitude/experience questionnaire. Administrative record information will be available prior to mathematics and science assessment and will include academic records, standardized test results and microenvironment description. For each student this will be augmented with parent and school interview data.

The research database created from HSLS-09 will constitute a national resource available to researchers for the study of the American educational system and of American students. The specific drivers for HSLS-09 are to characterize decision-making about post-secondary education and about persistence and attainment of credentials in STEM career trajectories and to track students' knowledge growth, both depth and breadth, in STEM areas.

The matrix structure of HSLS-09 must allow examination of both cross-sectional and longitudinal questions. One specific purpose of HSLS-09 is to provide direct information about knowledge increments and about individual changes in interests or expectations about post-secondary plans. It also must provide a data base for cross-sectional studies, for example linking assessment performance with attitudinal information or identification of barriers, or for examining issues such as parent-student [dis]agreement about expectations or expectation-course grade/performance inconsistencies either for the general student population or for significant subgroups of sufficient size. Adaptive tests have the advantage in the HSLS-09 context that they outperform conventional tests in the measurement of individual change (Kang and Weiss, 2008).

SECTION II

§ 2.1 COMPUTER-BASED AND COMPUTER-ADAPTIVE TESTING

With every half-decade younger cohort of students, comfort and even dependence on electronic technologies increases dramatically. By 2009, high-school students can be expected to be both facile and at ease with computer technologies that frustrate many of their mentors, making it possible to take advantage of advanced technological solutions for assessments and surveys of this population. Technological opportunities fall into two broad categories: data acquisition and organization, and item design or selection. Both of these are discussed in this report.

When the purpose of the test design is to assign a limited number of items or tasks to students in an effort to evaluate proficiency under time constraints, then CAT systems are essential. These tests assign items/tasks dependent on branching systems programmed within the computer and based on probability estimates of both students' proficiency levels as well as item/task parameter characteristics (*e.g.*, item difficulties, item discrimination indices).

In her 2007 Presidential Address to the American Educational Research Association, Eva Baker shared a vision for the future of testing with significantly more use of technology-based assessments. The implication is that the potential for assessment (CAT and otherwise) is as yet largely untapped pending the incorporation of continuing advances in item and test design.

The task force considered item and task formats that should be included within a computer-based or computer-adaptive testing (CAT) system. An example of an Evidence Centered Design (ECD, see § 3.1) complex problem-solving simulation in computer network troubleshooting has been described (Williamson, Bauer, Steinberg, Mislevy, Behrens and DeMark. 2004). These types of simulation exercises are ideally suited for computer-based testing platforms. Programming features of the computer can adapt different types of exercises to the estimated skill levels of students and data can be stored effectively as students work through steps of the problem; Dawson and Wilson (2004) give an example in text comprehension. Another desirable feature of computer-based systems is that steps students take toward

completion of their exercises can be time-stamped. As a result, computer-based systems of assessment not only can measure accuracy/inaccuracy of students' responses and the various strategies they employ during completion of tasks, but also the speed of processing at which examinees complete items or tasks can be recorded. Historically, both accuracy and efficiency have been hallmarks of expert performance in many domains (Keating, 1990). Computer-based testing systems are also ideal platforms for presenting scientific and mathematical visualizations (*e.g.*, animations, video-audio clips, graphical displays) as well as affording students' opportunities to manipulate or interact with these visualizations.

After careful deliberation, the Task Force reached a solidly affirmative position on the transition to CAT technology for longitudinal studies and identified potential strengths and advantages over previous assessment modes (both adaptive paper-and-pencil, and CBT). Discussions of the complexities of CAT covered issues of item pools, item selection, and sequencing as well as technical hardware and software considerations. Crucial discussions focused on the careful planning required to successfully implement CAT (or CBT): the near-heroic amount of technical development and coordination among the many disparate players, caution about excessive optimism *vis a vis* reuse of items from existing item pools and the time required to develop new items, the need for extensive piloting with the opportunity for near real-time revising, and good luck.

§ 2.2 CONSIDERATIONS AND REQUIREMENTS FOR CAT

What is important for each particular version of CAT (or CBT), is that the development be integrated with the development of the assessment - from the objectives through the constructs to the item types and the individual items themselves. The hardware and software possibilities in turn open up possibilities for constructs and items; while the specific decisions serve at the same time to limit the measurable constructs and the types of items. Planned in concert, CBT/CAT assessments can take advantage of the synergies among algorithms potential for complexity, hardware/software options and the varied natures of content and query types.

CBT and CAT require extensive upfront development. Fundamental decisions during the early stages of technological development/implementation include platform(s), specific standards, computer languages, database structures, modes for test transmission, and data transfer. These decisions are not independent of the study objectives or the constructs to be measured or the types of items to be included or the planned analyses. Fortunately, the co-development of the technological aspects of CBT/CAT with the assessment aspects can pay great rewards when: i) objectives inform the data structures and planned analyses, ii) constructs inform the adaptive algorithms, iii) hard/software capabilities allow/suggest new item types or responses, iv) computing flexibility leads to new partitioning/interspersing/interaction of sections of the assessment, v) flexible algorithms allow "on the fly" item sequencing to follow constructs or processes, vi) computing brings new opportunities for observing the student's processes of test-taking or for intervening.

As a practical matter, it should be verified that each student has the necessary IT skills to navigate through the test, although this is a rapidly diminishing problem for current and future cohorts. More importantly, it should be required that the test administrators have the necessary IT skills to trouble-shoot any hardware/software glitches that might occur during test administration.

Table 1. Advantages and Difficulties of Computer Tests

	PROs	CONs
COST	Primary Development Costs Low Costs for Repeat Administration Development Costs Replacing Labor Costs	Up-Front Costs: Development of Software Hardware Acquisition Large [CAT] Pretested Item Pool and/or Item Adaptation and Calibration
TEST MONITORING	On-line Monitoring and Addressing: Technical Difficulties Inattention-Random Responses	Test Administrators Must Trouble-shoot (*e.g.*, "frozen computer")
STUDENT RESPONSE	Opportunity to "Maximize Engagement" Opportunity to Deliver Feedback	Variation in Familiarity with Test Setting Variation in Hardware Critical Usability Issues for Hardware
DATA ISSUES	Instant Data Opportunities for Test Structure: Incorporate Prior Information Adjust "Level" Intra-Test Partial Scoring for "Level-setting" Data Quality: Real-time Verification Detection of Unexpected Performance Opportunity to Correct/Prevent Data Loss	No "Paper Trail" of Students' Responses Vulnerability to Truly Large-scale Data Loss or Theft (computer security & privacy issues)
WEB-BASED TESTS	Wide Availability Simultaneous Testing to Minimize Cheating	Web Filter Interference Difficulties Scaling-Up from Field Tests Test Administrators Must Trouble-shoot Challenge to Assure High Security

§ 2.3 TECHNOLOGICAL ASPECTS OF CAT

Hardware and Administration Conditions

Longitudinal surveys typically can be characterized by infrequent administrations, each generating a large volume of data acquired in a relatively short period of time. These conditions are not conducive to computerized administration of fairly secure and moderate-to-high stakes tests. If the testing period is limited by weeks instead of months, demand for equipment might be fairly high and would limit hardware choice to existing in-school computers in order to contain the cost or alternatively to shift to multi-day administration, or some combination of both.

It is imperative to clearly state the minimum standards for delivery of a CAT; OS, CPU, memory, speed, storage, type of optical driver, internet connectivity, video capability, screen size, portability, and backup options are among the variables requiring specification. The standard should be decided, insofar as is possible, based on the future availability at the time of assessment, and actual CAT design. Information on the hardware available in public and private schools may be known already. It should be expected that hardware is constantly evolving and is a moving target. A simpler choice to assure comparability and suitability of hardware is to provide a single standard configuration for use during the assessment process. This could also minimize the ancillary difficulties of hardware changeover between the 9th and 11th grade assessments.

If conventional multiple-choice tests of mathematics and science are used, the combination of screen size and resolution may not be an issue. However, if relatively lengthy descriptions and graphic information are necessary to set up the questions, screen size may become critical in determining the hardware to be utilized. Also, some simulations and other higher order skills may require manipulation of large amounts of information, making the combination of screen size and resolution critical. This seems to be an opportunity to introduce CAT-specific capabilities into the items, such as interactive items to investigate the rationale and thought sequences of word problem solving.

There are many test item delivery options; over the internet, CD/DVD optical disks, wireless within a class, or hard disk connected through local area network. Each option has advantages and disadvantages (see Table 3). To use CAT in a longitudinal study each of these can be considered, especially in regard to the projected administrative conditions. Delivery over the internet has advantages in terms of real time monitoring of administration progress, updating of testing materials, real time monitoring of item functionality and data collection, which might be especially useful for on- going testing. However, it has the very significant disadvantages of requiring a wider range of hardware to be accommodated, making control of display specifications and cost more difficult. Also, during internet-based testing it is necessary to monitor and control access to the general Web access in order to restrict searching for an answer on the web or doing some unrelated tasks instead of focusing on the test. For limited time survey assessment like HSLS-09, this potential hazard alone militates against internet testing in favor of the CD/DVD option with data collection LAN.

Administrative procedural environment in schools may vary great deal, for example other activities going on in the testing location while testing is in progress. Reducing this physical variability through operational standardization is sensible.

Software Issues and Decisions

There are many authoring computer languages. Delivery and administrative constraints may provide a good guide to selection of a suitable language for development and maintenance.

Modularize the software by functions during development, for example: interface, handling item presentation, items themselves with statistics, data capture, calculating/evaluation of responses, data base update, and reporting to administrator.

Since there will be no paper trail, it is essential to build in redundancy and continuous backup at every stage.

Database Questions to Answer First

The locus for monitoring the CAT activities needs to be clear at the outset; and the design of database should reflect the projected analyses using this data. A few relevant questions are:

1) Where are item banks stored?
2) How are banks secured?
3) How are data files secured?
4) What backup procedures need to be put into place to insure against bank and data loss?
5) Where are the data stored at various stages of administration: before test, during test, after test, and after all tests at a single site?

6) Who is going to monitor the test administration centrally and who will provide direct oversight of examinees?
7) Who should have access to data? Levels of access must be decided in terms of security and usefulness.

Computer Aspects of Scoring

Multiple-choice items provide the easiest scoring option for CAT; however, alternatives pose little or no problem for computer scoring. For example, short answer open-ended questions should be considered because these items often provide substantially more information than typical multiple-choice items without increasing testing time per item. Scoring of short answer open-ended questions can be carried out by computer without loss of information and without impediment to the assessment process. Furthermore, there is a great and growing array of types of computer-based task formats that can be increasingly open-ended from the point of view of the student taking the test, yet still allow automated scoring. Scalise and Gifford (2006) provide a useful taxonomy of "constrained response" tasks for CBT.

Consideration should also be given to other CBT/CAT possibilities. With the computer application it is possible to record every key stroke, latency, and all administrative conditions of stopping and pausing, opening up opportunities simultaneously to monitor or to prompt attentiveness: How long each item was displayed before any answer was made? How many times and in what sequence was the answer changed? This does not mean that everything that can be recorded is useful. By planning early using the measurement model, it is possible to identify the useful information that relates well to the outcome variables. Latency and keystroke history can be very useful in discriminating between thoughtful responses and random ones. Considering power vs. time as a characteristic of testing, with power being dominant, limiting scoring to only "correct," "incorrect," "omitted," and "not reached" might be the minimum-scoring criterion for reporting, but not the most powerful. Other variables may also be considered to examine test-taking behaviors that correlate with respondents' characteristics.

In a realistic testing context, returning to previous items is commonly practiced on a paper-and-pencil test as well as many short answer questions in the CAT context. Adaptive testlet design can allow changing the answer within the testlet, a significant advantage over an item-level adaptive test.

Polychotomous item models should be evaluated for difficulty on all response options including correct and wrong options of multiple-choice items by race, language and other additional variables of interest. Because of the multiple response alternatives, data required to estimate the parameters are much greater for polychotomous model.

Instrument Design for Computer Administration

Currently only a delivery mode change from paper-and-pencil to CAT has been seriously discussed. In this case the notable change of modes would only be the navigational procedures; and the amount of information on the screen would be most critical. However, the e-learning environment is very rich in media format including visual and sound information. It seems odd to have only text and graph testing items when much of learning is multi-media. If computer is going to be used for testing, one might as well take advantage of other capabilities as well. CAT can administer more interactive items instead of simply delivering paper-and-pencil items on a computer, with or without open-ended questions. The growing

realistic e-learning context may well have much to contribute to valid testing of learning beyond the current test paradigms including current implementations of CBT/CAT.

Scrolling (horizontal or vertical) is sometimes difficult, and students do often not notice hidden information. (However, agility in manipulating electronic information is itself a relevant, and measurable, skill when it contributes to information identification, selection and evaluation.) For conventional test items, this potential problem is easily resolved. Scrolling can be eliminated if test designers provide a button for students to click to see the next/previous page so that they reconceptualize long items as "multi-page" item - like an electronic book.

CAT does not have to entail domain-at-a-time administration. It is possible to mix items across domains and also to use covariance information among skills to adapt measures for several skills. Whether domains should be mixed within a test is an empirical question that has not yet been answered adequately. One should certainly consider using information from other domains to inform succeeding item or testlet selection.

A caveat to adaptation of pencil-and-paper items written first in one language then translated is that translation can also alter the level of difficulty; so, level of difficulty needs reevaluation for the translated version. Alternatively, items can be developed in multilingual forms simultaneously if multiple language versions are necessary; however, it is very difficult to establish comparability when both versions when these are evolving simultaneously.

To provide accommodation for persons with disabilities, the most efficient approach is to let this task follow the complete development of the assessment in its primary or predominant form, just as the more efficient approach to development of multilingual versions is to completely develop the items in computer format in the primary language, then return to produce translated versions.

It remains important to verify that each student has the necessary computer skills to navigate through the test. Appropriate instructional sequences are necessary to ensure that the student knows how to answer the questions and can easily perform all the computer tasks needed.

Field Testing

For CAT field tests should be carried out which will provide sufficient basis for assembling multistage test forms. Existing parameters for items available from previous projects cannot be taken at face value, as they were estimated on different scales, with different populations, using different models. Successful field-testing involves use of a fully representative sample of examinees, use of a sufficient number of items and use of the same testing mode as envisaged for operational testing (*e.g.*, on notebook computers, web-based delivery, etc.). Item parameters can be refined after the operational data are in hand. As a rule of thumb, including 1.5-2.0 times as many items in the field test as are anticipated for the final form pool is recommended in order to allow for culling poorly performing items and flexibility in form construction. A sample of, say, 750 persons for each item is desirable for estimating item parameters (see § 3.1 and § 3.2).

In general, an approximately representative sample is extremely desirable for a field test rather than a sample of convenience. In addition, supplemental selective over-sampling of uncommon subpopulations of particular interest may be advisable. The concern is to achieve approximately equal performance of the assessment over the full range of the scale and across all subpopulations of interest. To implement

the designs and analytic paradigms presented in the next section, approximations to the parameters associated with individual items and with their joint distribution are required. Thus, in order to assemble forms for adaptive multi-stage testing, at least approximations to the relationships among various domain (*e.g.*, mathematics and science) proficiencies and student-background data (previously available and/or gathered during the assessment) are needed. (In § 3.1 and § 3.2 of this report, these relationships are expressed mathematically as the joint distribution of $\mathbf{Y}_B$, $\mathbf{Y}_S$, θ_M, and θ_S.) Precise estimates of item and distribution parameters are not required for assembling forms, again because they can be refined from operational data later.

Post-hoc simulation is a useful technique for evaluating/confirming the effectiveness of a particular configuration of CAT (Weiss and Gibbons, 2007). When an item bank has been developed and calibrated to implement CAT, post-hoc simulation based on actual responses can investigate features of a CAT: ways of structuring the item bank, numbers of items to be administered by stage, various scoring methods. Implementation requires item responses on all items in the item bank by a group of examinees (about 200 should suffice). Post-hoc simulation then "readministers" the items adaptively according to the CAT scheme and any variants under consideration to compare the scores under these options with the scores from the entire data bank. These data are then analyzed to identify the CAT configuration that maximizes the test performance or to find the overall efficiency of the selected CAT configuration with the much longer conventional test design.

§ 2A IMPLICATIONS OF CAT 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. Under these constraints, adaptive testing appears to be the best way to measure the constructs of interest; the advantage of CAT over CBT is to focus questions rapidly enough to probe knowledge depth or identify the specific elements critical to decisions.

The two planned assessments (early 9th and late 11th grades) can be done using a single platform and supporting software to eliminate hardware/software compatibility and comparability difficulties. 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 the first stage of each testlet (see *2.5 Stage Designs* in § 3,3).

CAT offers specific advantages with regard to utilization of additional information because *adaptation* is equally relevant to the questionnaire construction; and inclusion of external information as a basis for adaptive sequences of items is even more efficient in the background survey of interest/attitude. For example, this survey content of HSLS-09 focuses on opportunities either to introduce or to alter policies. The longitudinal structure of HSLS-09, coupled with CAT capability for real-time comparison of a student's response in 11th grade with the earlier 9th grade response, can be used to pinpoint the timing as well as the target for interventions whether with i) the student, ii) the family, iii) the school *vis a vis* student-oriented programs, iv) the school *vis a vis* teacher qualification, performance, training opportunities or v) the school *vis a vis* staffing, resource allocation or administration.

The HSLS-09 database can be expected to be used (longitudinally) for modeling student trajectories and cross-sectionally for comparing different student populations both with regard to academic growth and with regard to interest in post-secondary education. Consequently, both longitudinal and cross-sectional analyses will require comparable precision over most of the testing scale. Appropriately designed, CAT is well suited to such broad assessments with comparable precision requirements over a fairly wide range of performance because of the gains in scoring precision for individual students.

Effective use of the technological opportunities should enable HSLS-09 to go beyond simply computerizing paper survey instruments; but this will require tapping IT expertise in the examination of the goals of HSLS-09 to ascertain which of these can be streamlined, and which can be enabled with existing, already proven technologies. For example, it may be possible via technological means to increase inter-assessment information to capture promptly the dynamics and rationales for decisions students reach between assessments. Or, it may be possible to motivate continuing student [or parent or school] participation through feedback about participation, about the aggregate of participating students, or individually to and about the student him/herself.

Adaptive Test Structures for HSLS-09

Neither of the adaptive testing strategies widely utilized in other implementations, two-stage split test and item-by-item adaptation, is well-suited to HSLS-09 due to the time, and hence length, constraint. The two-stage split test would require too many items in the first stage to avoid an unacceptable frequency of misrouting. The item pool requirements for item-by-item adaptive test would make that structure prohibitively costly in both calibration time and money. For HSLS-09, efficient testing can be accomplished with a testlet-adaptive design that reduces the number of contexts to a few, with several items (*i.e.*, testlet stage or complete testlet) addressing a particular concept for each context. In the case of science, perhaps 3-5 contexts in environmental science could serve as the source for items in general as well as domain-specific sciences. In the case of mathematics, the natural division might be according to topic areas or mathematical concepts.

Of the various types of adaptive designs, testlet-adaptive designs (in particular see **§ 3.3**, *2.5 Stage Designs*) are best suited to HSLS-09. These allow intelligent choice of an inaugural item level, followed by adaptation. Further adaptation occurs within each testlet, allowing several opportunities for correction of an inaccurate initial choice or of chance correct responses. This approach also allows a subdivision of questions on science between general scientific reasoning and more domain-specific items, ideally from a single contextual setting. To the extent possible, use of partial scoring at the final (second) stage is desirable to increase precision. Thus, for example, by setting a context in environmental science, the initial testlet sequence could address general scientific reasoning; a second testlet sequence could be domain specific - biology, chemistry or physics, with general science as the default. Such an approach would allow longitudinal comparison of growth in a student's area of scientific strength and/or interest as well as cross-sectional analysis of students by population subgroup, interest, science domain, or other feature likely to be predictive of persistence and attainment in STEM majors and careers. Even with an astute choice of specific design and adaptive algorithm, the precision attainable by HSLS-09 will be constrained by the total time available for test administration.

The challenge is to find a rapid, reliable basis for determining a starting point to probe the depth of a student's understanding about a particular concept without losing the potential to re-adjust the level.

Fortunately, since students need only be rather broadly classified for this purpose, available external information - later coupled with internal information (*e.g.*, responses to earlier test items) - should be adequate. (External information could include history of courses taken and grades in those courses, with additional information for the 11th grade test: additional courses, scores and even the item responses to the 9th grade test, and scores on broad-based state or national tests (*e.g.*, PSAT, NAEP, AP) were these available.)

Implementation Decisions for HSLS-09

For HSLS-09 the advantages of CAT are chiefly: 1) the potential improvement in scoring precision for each student from using external information to inaugurate the sequence of item selection, 2) the greater range of item formats and 3) the improvement in data capture and data handling. Achieving these advantages requires addressing several specific technological issues.

- Where are the data to be stored at various stages of administration, before test, during test, after test, and after all tests in a school? How will data be backed-up securely at each stage?
- Who is going to monitor the test administration centrally and who will be responsible for direct oversight of the students? An administrator [contractor staff] may be interested in this control.
- Who should have access to data? Levels of access must be decided in terms of security and usefulness.

SECTION III

§ 3.1 ASSESSMENT DESIGN

Design of an assessment should consider future analyses in gathering data. For example, it seems clear that one use of the results of HSLS-09 would be to construct a propensity score (Rosenbaum and Rubin, 1983) based on all of the independent variables with the binary variable 'STEM or not' as the outcome; then the propensity score could be used as a matching variable to estimate the effects of various 'treatments'. If so, then the data ought to be collected and reported in ways that facilitate such analyses. Although it might be a bit premature in the early planning phase, a discussion of the format of the data presentation ought to take place, both as part of the CAT development specifications and as clarification of the principal study objectives, ensuring that these will be met successfully. See Wainer (2000, 2005) for an extensive discussion and illustration of how this might be done well and how it can be done poorly.

To the extent possible, the actual planning of a longitudinal study will benefit by following the stepwise paradigm for evidence-centered-design (ECD) and by considering the primary communications (analyses) that the study must provide.

Evidence Centered Design

Evidence-centered assessment design (ECD) provides concepts, processes, and representational forms to guide the work of task developers (Mislevy, Steinberg, and Almond, 2003; Mislevy and Haertel, 2006). ECD views an assessment as providing the basis for an evidentiary argument: reasoning from what we observe students say and/or do in a few particular circumstances, to what they know or can do more broadly. It makes the underlying argument more explicit and operational elements easier to share and reuse.

Evidence-centered design applies the concept of layers to the processes of designing, implementing, and delivering an educational assessment. Thus, following ECD yields a documentary trail to establish validity. Each layer has its own key concepts and entities, and knowledge representations and tools that assist in achieving each layer's purpose. The five layers can be briefly described as follows.

Domain Analysis gathers information about the domain to be assessed. If the assessment being designed were to measure science inquiry at the elementary level, domain analysis would pull together concepts, terminology, representational forms, and ways of using them.

Domain Modeling organizes information and relationships discovered in Domain Analysis along the lines of assessment arguments. Domain experts, teachers, and designers work together here to lay out what an assessment is meant to measure, and how it will do so. For example, domain modeling responds to a query such as: Just what kind of knowledge is important *vis-à-vis* [high school] science inquiry, and what assessment situations best allow students to demonstrate this?

The Conceptual Assessment Framework concerns technical specifications for the machinery that constitutes an assessment, such as measurement models, scoring methods, and delivery requirements. Data structures, scoring algorithms, and measurement models can be used not only for science inquiry, but re-used in assessments in other areas and for different purposes.

Assessment Implementation includes authoring tasks, calibrating items, finalizing rubrics, producing materials, producing presentation environments, and training interviewers and scorers, in accordance with the assessment arguments and test specifications. *Assessment Delivery* concerns presenting tasks to examinees, evaluating performances, and reporting the results.

Working through the ECD layers helps ensure a coherent design that is tuned to the purpose of the assessment.

§ 3.2 FULL BAYESIAN ANALYSIS

The statistical framework discussed at the Task Force meeting to support CAT for longitudinal surveys was that of item response theory (IRT) under a fully Bayesian paradigm. Most of the discussion at the meeting concerned the case in which one-dimensional IRT scales would be appropriate for characterizing the constructs of interest, and enough items would be administered per person to support proficiency estimates at the level of individuals.

IRT provides the basis for administering different test forms to different students, even adaptively, yet obtaining results on common scales. The machinery of IRT can be used to construct two-stage tests with first and second stage forms. Partial-credit and nominal response IRT models can be used to increase precision of estimates for students by capitalizing on additional information that may be available in non-correct responses (Bock, 1997; Muraki, 1997; Samejima, 1997), and testlet IRT models can be used to handle conditionally dependent responses within testlets (Wainer, Bradlow, and Wang, 2007). The Bayesian framework provides a framework for sequential testing and parameter estimation with complex IRT models from sparse data (at the level of subjects), utilizing joint information across scales, and producing finite scores for all response patterns.

The Case of Unidimensional Scales

Using IRT, results can be obtained on common scales when the assessment, whether fixed or adaptive multi-stage, utilizes different test forms for different students. This same paradigm can be used for surveys in which estimates for individuals are required, as in HSLS-09, or those in which only population estimates are needed, as in NAEP.

Given the constraints of severely limited testing time and the production of scores for individuals, univariate IRT models are currently envisaged for Mathematics and Science in HSLS-09; the latent variables to be measured will be operationally defined through the selection or creation of tasks. Denote these latent variables by θ_M and θ_S, which can further be subscripted by *i* for students when required. Given the breath of content in both math and science, even if the focus is on reasoning rather than details of content, both domains are surely *not* strictly unidimensional. The unidimensional model is viewed as an engineering approximation - a mechanism for mapping performances on forms of differing difficulties onto the same approximate metric. Inferences would be limited to the collection of items used in the surveys, rather than attempting to define a learning domain or proficiency scales more broadly. Inferences beyond this collection of items, *i.e.*, to other tasks, scales, or proficiency definitions, would need to be supported by additional studies.

Unidimensional scales for "general" math and science (again, however defined by item selection) were envisaged for Grade 9. For Grade 11, two alternatives were proposed. One was maintaining these scales, test forms and adaptivity schemes, and interpretations, in both math and science. The other alternative did this in math, but differed for science. It would administer different 11th graders different, more content-specific, forms based on background variables - either course-taking or personal choice - and providing results for Grade 11 science based on performance in a selected (self-selected if choice is permitted), presumably maximal, subarea. This approach violates the conditional independence & unidimensionality assumption of the IRT model as a measure of *global* science knowledge, but nonetheless could support an interpretation of the scores as performance biased toward students' areas of strength. Problems with interpreting performance on self-selected (specifically, not missing at random) tasks under IRT are discussed in Bradlow and Thomas (1998), Mislevy and Wu (1996), and Wang, Wainer, and Thissen (1995).

More ambitious models could be fit to the data, including for example mixtures of IRT models among latent classes of students with different profiles of background knowledge and insight. Such models could address hypotheses about aspects of proficiency that previous research has suggested may be related to STEM choices, as per Prof. Kulikowich. The time limitation won't provide enough information to fit such models and estimate scores for all individuals with any accuracy; these hypotheses would need to be explored in terms of population structures such as classes, conditional item parameters, etc. Multistage tests would make these analyses more challenging, although the missing responses would in this case be missing at random. For these reasons, such analyses should be considered as possibly fruitful options for secondary research, rather than as the main ways for constructing forms, analyzing data, and summarizing results.

Some Details

Denote observed response vectors for math and science respectively by $\mathbf{X}_M$ and $\mathbf{X}_S$. Response vectors can be partitioned into those resulting from first stage and second stage test forms, *e.g.*, $\mathbf{X}_M = (\mathbf{X}_{M1}, \mathbf{X}_{M2})$. Denote student covariates by $\mathbf{Y} = (\mathbf{Y}_B, \mathbf{Y}_S)$, where YB refers to those ascertained by transcripts and other records and YS refers to those from the student survey administered along with the math and science tasks.

The number of testing stages discussed at the meeting was two, with about three-to-four possible first stage forms, and routing to about three second stage forms after each first stage form. In some cases, a given second stage form could be reached from different first stage forms.

The construction of the forms would be based on a partitioning of the population expected to be adapted from the unselected population for Stage 1; and for Stage 2, the subpopulation who had been routed to a given Stage 1 form. A criterion such as equal numbers of students being assigned to each of the next stage form options can be employed to construct forms. The construction of the forms would seek relatively flat IRT test information curves across the expected group of students to be administered the form.

Luecht and Nungester (1998, 2000) describe methods to assemble multi-stage tests in their computer-adaptive sequential testing (CAST) system, and Edwards and Thissen (2004, 2007) describe a computationally-intensive method to design multi-stage tests with uniform item exposure (uMFS systems). Either CAST-like or uMFS-like systems could use background information to select the first-stage test (which is a block of items, sometimes referred to as a *testlet*) and would provide branching rules to the second-stage testlets based on the observed first-state outcome.

Information from $\mathbf{Y}_B$ and/or $\mathbf{Y}_S$ can be used to select the first stage of the first test, say Math; that is, selection for Person *i* with covariates $\mathbf{Y}_i$ is based on $p(\theta_M | \mathbf{Y}_i)$. Performance on the first stage test, $\mathbf{X}_{M1}$, and again perhaps $\mathbf{Y}_B$ and/or $\mathbf{Y}_S$ would be used to select the second stage form of this test; that is, $p(\theta_M | \mathbf{Y}_i, \mathbf{X}_{M1i})$. Information from the results of the Math test and again perhaps YB and/or YS would be used to select the first stage for the Science test; that is, $p(\theta_S | \mathbf{Y}_i, \mathbf{X}_{Mi})$. As seen for example in the uMFS system of Edwards and Thissen (2004, 2007), these routing rules can be fairly simple, based on a simple summary of the background information and then the summed score on the first stage test, irrespective of IRT and estimation procedures that will be used later.

These routing decisions can be pre-computed, through IRT, based on simple functions of responses on stages, *e.g.*, total stage scores. In this case, complex IRT estimation need not be carried out in the field, during administration, for each student. The loss of precision is minimal, given the crude adaptivity decision to be made, and robustness in the field is desirable (Luecht and Nungester, 1998, 2000; Edwards and Thissen, 2004, 2007). More complex IRT estimation of individuals' scores can be carried out when all the data are in hand and clean, and IRT item parameters have been refined. The full Bayesian model uses pattern scoring of all responses each person has provided, exploits the correlation across scales, and produces a posterior mean vector and covariance matrix for each person. Thissen, Nelson, and Swygert (2001) describe several alternative scoring systems that are based on more or less of the response pattern - one can use the item response pattern, the pattern of summed scores across stages, or the total score over all stages adjusted to equate across various paths through the multi-stage test (Stocking, 1996).

§ 3.3 ASSESSMENT EFFICIENCIES AND SCORING

The need to streamline assessments in response to severe time constraints and the need to balance content or context are justifications for constructing an assessment from testlets. Each testlet is a "group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow." (Wainer and Kiely, 1987). For items with substantial preliminary exposition or "context set-up," reuse of the contextual information allows either greater detail or a larger number of items relative to the time spend reading and absorbing the context. An entire test then consists of several testlet sets, each as described below with the entry performance level recalculated for each testlet set from background information and performance of preceding sets. The number of testlet sets and the length of each testlet stage will depend on the available time and the scope of the content to be assessed.

Psychometric advantages include treating each testlet as a test in miniature; this can serve also to minimize undesirable context effects between items and item sequencing effects. In an adaptive testing setting, adaptation becomes more flexible and can be structured to occur within each single testlet either independently or dependently across testlets.

Scoring Testlets

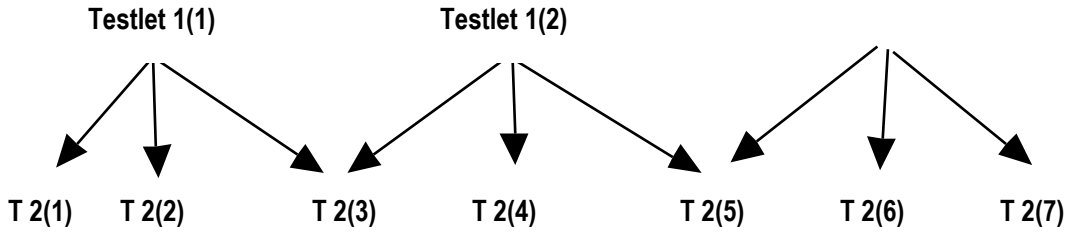
There are many options, and these are laid out in detail in Wainer, Bradlow and Wang (2007), hereafter WB&W, and can be done with all sorts of easily available software - although the full Bayesian model with covariates can only be done with SCORIGHT (3.1) (Wang, Bradlow and Wainer, 2004). See also Thissen and Wainer (2001), hereafter T&W, for expanded discussion of a range of test scoring strategies.

Two options were discussed in some detail at the first Task Force meeting, specifically scoring a testlet as a polychotomous item, a k-item testlet that yields a single polychotomous score from 0 to k, or as k items with possible local dependence. The former approach is quick and easy, but because it collapses all response patterns with the same number right into the same category it can potentially lose information (see figure 10.2 in WB&W, pages 150-151 for details). A second disadvantage of the polychotomous scoring approach is not of immediate concern, but may be important in the future, is that it does not allow testlets to be made up adaptively on-the-fly. For example, there might be one figure and associated with it 15 or 20 items, but the idea is to ask only 5 or 6 items to each examinee, determining which 5 or 6 will adaptively as a function of the examinee's responses. Obviously, it would not be possible to calibrate in advance all of the 15,504 (20 choose 5) 5-item polychotomous items that might emerge. But the full testlet model could accommodate this strategy easily. Using the fully Bayesian testlet scoring model will require some cleverness of application (pre-calibrating the testlets to estimate the amount of local dependence).

Adaptivity for Two-Phase Testlets

The background information should be used to set the initial level ("prior") for each examinee. On the basis of that prior one of several (say three) different initial testlets are assigned. There may be some overlap in items among these phase-one testlets. The examinee's posterior distribution calculated after the initial testlet will determine which phase two testlet is assigned. Shown below is an illustration of such a scheme in which there are 7 phase two testlets. Note that there is a possibility that an examinee

initially routed through testlet 1(1) can end up with the same phase two testlet as someone initially assigned testlet 1(2) - that is, the data can, to some extent, over-ride the prior.



Efficiency Gained by Scoring Partial Knowledge

A polychotomous model (*e.g.*, Bock's nominal model, 1972) can be used to provide a small increase in information within item. So instead of scoring an item correct/incorrect partial credit is given depending on the particular incorrect answer that was chosen (Thissen, Steinberg and Fitzpatrick, 1989).

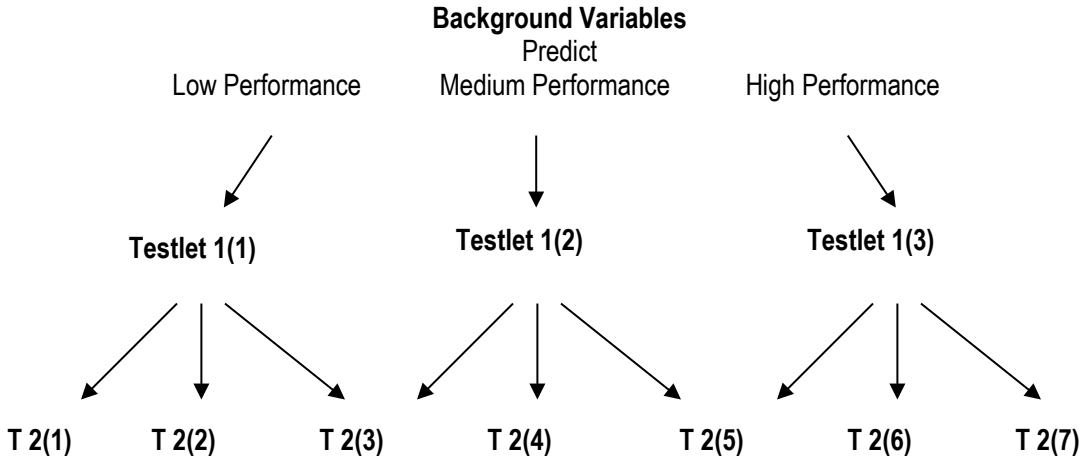
The strategies mentioned above are especially important in this application since the goal is to extract the maximal information in the least time. Thus, the extra computational work in (i) choosing empirical priors carefully, (ii) allowing different within-testlet response patterns to yield different scores, and (iii) getting information from distractors can yield a much-needed dividend of extra precision.

Using one score (say math) to aid in the calculation of the prior for the other score (say science) is important (see chapter 9 in T&W). This, when combined with the technology discussed above can make a short test as accurate as an unaugmented much longer one. It cannot however help in spanning a broader range of content or of difficulty. To span the largest amount of content possible in the time frame allocated, some version of the multiple-choice format is the only candidate. 'Some version' is taken to mean that there are very attractive alternatives to the traditional item stem and five choices. For example, have a list of item stems (say 1- 20) and associated with them another list of answers (say A-Z), in which the examinee's task is to adjoin a letter response to each stem, where response choices can be used once, more than once, or not at all. By reducing the likelihood of a successful guessing strategy such a format boosts the information in each response.

In a realistic testing context, returning to previous items is commonly practiced on a paper-and-pencil test as well as many short answer questions in the CAT context. Adaptive testlet design can allow changing the answer within the testlet, unlike an item level adaptive test.

2.5–Stage Design

This test design, discussed in some detail by the Task Force, exploits the idea that student background data (or "covariates") Y might serve as the basis for selection of a routing test, or the first testlet, in a conventional (Lord, 1980, chapter 9) two-stage adaptive test. A graphic depiction of such an adaptive test is shown below.



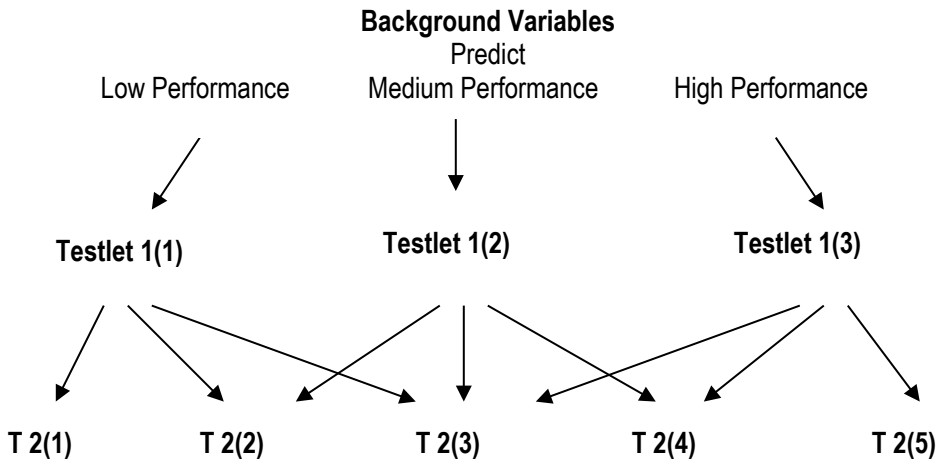
In this (arbitrary) depiction, the student background data (or "covariates") **Y** are used to "route" students into one of three first-stage testlets [1(1), 1(2), or 1(3)] which are designed for students of relatively low, medium, or relatively high proficiency respectively. The responses of the students on the first-stage testlet, $\mathbf{X}_{m1}$, possibly in concert with **Y**, are then used to route the student to one of a more finely graded set of second-stage testlets. In the graphic above, there are seven second-stage testlets, ranging from very low to very high in difficulty.

Unlike a conventional two-stage test, this design has two points of adaptation: The first is between the collection of the background data **Y** and the second between the first and second stages of the test itself. It is not, however, strictly speaking, a three-stage test (although it has as many points of adaptation as a three-stage test). Hence, a name for this might be a "two-and-a-half stage test" or a "two-plus stage test."

Two points of adaptation are far superior to the one that comes from the use of a two-stage test alone, because the second adaptation can "correct" for routing errors that may arise if there are only two stages and one adaptive choice.

Scores on this "two-and-a-half stage test" or a "two-plus stage test" may be computed using any of a number of approaches based on item response theory (see Thissen and Wainer, 2001, chapters 3, 4, 7, and 8 for a summary of alternatives). Scores may be based solely on the item response data **X** or on the combination of **X** and **Y**. The choice between those options may depend on the use of the scores.

The graphic below shows a more realistic two-plus stage test design, in which there are only five second-stage testlets, ranging from easy to difficult. Note that, comparing this more realistic design to that above, there is more overlap between second stage testlets that can be reached from each pair of first stage testlets.



In the construction of multi-stage tests, it is neither practical nor desirable to construct testlets that measure only a very narrow range of proficiency, nor to route examinees extremely precisely based on short testlets. The result is that, in practice, the proficiency distributions of students routed to each of the three first stage testlets overlap substantially; so, at the second stage (of testing) those students are routed to overlapping testlets.

§ 3A CAT DESIGN AND ANALYSIS FOR HSLS-09

Logically, considerations of methodology should follow a clear explication of the measurement goals. In the case of HSLS-09, students' postsecondary and career trajectories are most surely related in a complex manner to (i) student and family characteristics, (ii) school characteristics and (iii) student achievement. Hence HSLS-09 will collect extensive background data to provide information on the first two sets of variables and some on the third. In regard to persistence and attainment in STEM majors and careers, the challenge is to measure as precisely as is possible those aspects of math and science achievement and interest that would be most useful in strengthening the predictive accuracy of the models to be constructed from the full data base to be built from HSLS-09. The degree of precision that will be attainable will depend on the efficiency of the methodological approach, particularly given the relatively small fraction of students expected to pursue STEM-related paths.

The key question for HSLS-09 remains: *What should be measured?* The answer depends whether the goal to create a single score or a profile. (Choices and some of the specific difficulties associated with adequately assessing STEM subjects are discussed § 4.A.)

More specifically, the question can be rephrased: What should be measured at each grade? Evaluation of growth becomes complicated as courses beyond the 9th grade level diverge with alternative tracks and accelerated sequences in mathematics and with differing sequences of science courses. Growth in mathematics for one student through a course in calculus may be accompanied by weakened memory of algebraic tricks or geometry theorems that would earlier have formed a natural basis for items. Similarly, in science, growth may be greater either in terms of broadening (*i.e.*, learning a different science) or in terms of deepening (*i.e.*, expanding knowledge within a given domain) or both. A viable solution for mathematics is to allow access to the full range of mathematics items at both 9th and 11th grade

assessments. This would permit the starting point (initial question at the first stage of a testlet) to be as close as possible to the student's *known* exposure/mastery of mathematics; and progress can be measured along the continuum of the entire range of items.

For broad-based longitudinal studies in general, one cogent argument for CAT is to use a broad item pool and a carefully chosen algorithm to construct a database of profiles as completely, as accurately and as efficiently as the available testing time allows. Many if not most users of this database who conduct secondary analyses will presume that the published estimates of individual students' scores are unbiased IRT estimates. Thus, in HSLS-09 it may be desirable to use only total population distributions for Grade 9 and for Grade 11 as priors for all individuals, so that the resulting estimates (obtained as described above) behave similarly to the unbiased IRT estimates.

CAT Design for HSLS-09

Ideally, the precepts of evidence-centered-design (ECD) for assessment will be applied for HSLS-09, although the imminent scheduling of the first assessment may not allow for adherence to detailed specifications. The first step is to identify the target constellation of constructs in order to follow the ECD steps.

Design of the assessment, despite the different performance expectations for grades 9 and 11, can take advantage of CAT by utilizing the full range of the (single) test instrument at both time points. Adaptive selection of the difficulty level for the inaugural test item at each time point can be based on external information with the expectation that students in the 11th grade would often start at a higher level. Such external information (see "Two and a half - stage designs") can be drawn from course history, grades, interest statement or (in 11th grade) prior results for HSLS-09 in 9th grade. This approach requires amassing a broad item pool, selection of external information and choosing an algorithm that yields a profile of aspects of math and science - or as much of a profile as those 40-50 minutes of testing allow!

Selection among possible CAT algorithms might differ at least marginally for the two grades since capability at the 9th grade would likely depend on (i) exposure to, and facility with, algebra; (ii) success in problems involving mathematical reasoning; (iii) analytic reasoning with texts incorporating scientific terms and concepts; presumably these same aspects would also help to predict future STEM course-taking and interest. In the 11th grade, stronger predictors of persistence and attainment in STEM trajectories might be expected to be (i) facility with discrete mathematics, as well as with statistics and probability; (ii) general science reasoning; (iii) appreciation of the structure of a particular science discipline. However, to the extent that distinctions between expected predictors for 9th and 11th grade depend heavily on attainment, a single adaptive algorithm constructed for the full-range test instrument would appear both feasible and desirable. This commonality of algorithm and use of a single test instrument preserves direct longitudinal comparisons at no expense to cross-sectional analyses.

Analyses, in particular of the connections between assessment performance and career decision-making, are likely to be complicated. It appears likely that linear regression or logistic regression models will not be up to the task. Rather, some non-linear or threshold models will be necessary. At this point, it is not clear whether this should play a role in the choice of CAT algorithms.

Implications of External Information on Students' Backgrounds

It is clear that with the small proportion of students that go on to STEM careers (3 to 8 percent were the estimates offered - possibly much smaller among some minorities and women) the power of the study to detect differences in independent variables will be small without either a gigantic increase in sample size or an over sampling of relevant individuals. The question is how to know in advance whom to over sample. This question can be answered cheaply and easily with a modest case-control study in which people in STEM occupations are the cases, and demographically matched people are the controls. To minimize age-cohort effects participants should be sampled who are as young as possible (24-25-year olds perhaps - the end age that the longitudinal survey anticipates). The "risk factors" that thus emerge will help in two ways, (i) over sampling participants into the study that have those factors, and (ii) being sure to include those factors as independent variables in the study.

This sort of preliminary study could be done with existing data (for example, AAMC has such information on students who take MCAT, as well as the information as to whether they go on to apply to medical school, whether they get in or whether they change from medicine to other STEM career.). Such a study, conducted over only a few months, could be invaluable in directing the longitudinal study and is likely to increase its potential usefulness substantially.

SECTION IV

§ 4.1 CONSTRUCTS

There are a variety of constructs that are important in understanding persistence in school and attainment of degrees and careers in STEM-related fields. Historically, these constructs have been primary contributors to competent and proficient learning and performance, and as such, are hallmarks of expertise (Alexander, 2003). Two important constructs which researchers are encouraged to study longitudinally from the middle- to high-school academic levels and throughout the collegiate and graduate years are principled knowledge (Alexander, Murphy, and Woods, 1996) and reasoning with scientific or mathematical information (Cognition and Technology Group at Vanderbilt [CGTV], 1990); Kulikowich and DeFranco, 2003; Wolfe and Goldman, 2005).

Principled Knowledge

Principled knowledge is more than the accumulation of vast stores of knowledge. This form of knowledge actually binds conceptual and procedural knowledge together around key ideas, theories, and practices of the domain (Rittle-Johnson and Star, 2007). Both experts and non-experts may be able to state or define principles. However, experts, or those who persist and attain careers in specific fields of study, tie significantly more knowledge elements, employ procedures, and execute practices around those principles than do non-experts or those who do not seek to acquire proficiency in a specific career or domain specialization. As a result, researchers have argued that principled knowledge changes qualitatively in time as learners have more opportunities to study a given domain (Alexander, 2003).

With respect to testing, Baker (2007) identified knowledge and use of principles in learning as one of the important qualifications to be assessed in future developments in assessment. Similarly, the National Assessment of Educational Progress (NAEP, 2007) recognized principled knowledge and the various forms

of reasoning that depend on principled knowledge essential in evaluating students' knowledge acquisition and understanding of scientific and mathematical information. Items in the NAEP database are built upon the premise that principled knowledge is essential in acquiring competence and proficiency in science and mathematics.

Scientific and Mathematical Reasoning

Reasoning, both text-based (*i.e.*, literacy, Wolfe and Goldman, 2005) and scientific or mathematical (Dunbar and Fugelsang, 2005) comes in many forms (*e.g.*, analogical, causal, deductive, inductive, and integrative) and has an extensive literature base (Thagard, 2005). Generally, forms of reasoning relate to cognitive processing mechanisms used during complex problem solving and reading comprehension tasks. In scientific fields, Chinn and Brewer (2001) argue that reasoning about data and methodology used to collect data is essential in scientific discovery, progress, and theory building. Likewise, in mathematics, reasoning with various forms of representation (*e.g.*, algorithms, spatial-graphical visualizations, verbal expressions) is key to acquiring proficiency in domains such as algebra and calculus and central to the mathematical modeling required in many scientific endeavors (Koedinger and Nathan, 2004; Wilensky and Reisman, 2006).

With respect to testing, several renowned scholars in education, cognitive psychology, and measurement (Baker, 2007; Mislevy, Steinberg, and Almond, 2003; Mislevy, Steinberg, Breyer, Almond, and Johnson, 2002) attest to the significance of designing measures and establishing the psychometric properties of scores for tasks that measure scientific and mathematical reasoning. Further, as with principled knowledge, organizations, such as NAEP, which develop tests and tasks according to curriculum standards, emphasize the salient role of scientific reasoning in monitoring developmental progress in science, mathematics, and related fields.

Constructs and Facets

Related to the ECD paradigm (Mislevy, Steinberg and Almond, 2003), three terms were used to describe item and tasks properties for the longitudinal CAT system. These terms were: a) constructs, b) facets, and c) covariates. (Covariates are variables measured outside of the CAT system that may be useful in establishing individual difference profiles of students that may be useful in guiding the initial branching of tests or items within the longitudinal CAT system. These covariates include variables like: a) grades in mathematics and science classes, b) interest in domains, and c) completion of specialized coursework such as Advanced Placement (AP) classes.)

Constructs are the latent traits measured within the CAT system, such as principled knowledge and scientific and mathematical reasoning. *Facets* are key processes or procedures that are evident when students demonstrate use of the primary constructs in learning and performance activities. For example, the facets evident in the use of science principles during problem solving as presented by NAEP (2007) might include: a) the explanation and prediction of observations when studying phenomena; b) provision of examples that illustrate the principles; and, c) the evaluation of alternative explanations based on principles that explain patterns of observation. Each of these facets can establish the framework for an ECD that allows scores to represent multiple aspects of the constructs. Mislevy, Steinberg, Breyer, Almond, and Johnson (2002) illustrated how the ECD paradigm can be used to measure problem solving with simulation-based assessments in dental hygiene. Examinees had to study patient cases simulated to

depict different dental hygiene problems. Problem solutions depended on examinees' information gathering and use, medical knowledge, and knowledge of ethics and legal matters. Evidence for each of these three constructs was manifest in several observable variables that arguably represent facets of the key constructs. For example, information gathering and use was defined and featured in the assessment system as: a) adapting to situational constraints, b) addressing chief complaints, c) evaluating examination procedures, d) evaluating patients' histories, and e) collecting essential information. The task force recommended that similar design strategies incorporating facets within the definitions of the primary constructs be considered when selecting or modifying items from existing item databases (*e.g.*, NAEP, NELS, PISA, TIMSS).

Multivariate Constructs, Undetermined Scales, or Very Short Test Forms

Most extant procedures for adaptive testing are for unidimensional scales. Item-level adaptivity in multidimensional IRT situations is discussed by Segall (1996) and Veldkamp and van der Linden (2002), but much less experience and theory is available to draw upon at present.

An alternative form of adaptivity for multidimensional scales that are at least moderately correlated is to adapt at the level of the test form, where test forms are heterogeneous with respect to scales but similar in levels of overall difficulty. This strategy is employed in the adaptive version of the PDQ literacy assessment described by Dr. Yamamoto. Decisions for sequential form selection can be made on the basis of, say, total score on a first-stage form, ignoring the subscale structure. This approach can be used with either multivariate IRT models or parallel unidimensional IRT models. The adaptivity desired here is simply to route examinees to an appropriate neighborhood of items, rather than precise maximization of information. This strategy can also be employed in the cases in which either final scaling decisions have not been made, multiple alternative models might be fit, or the test length is too short to estimate individual-level proficiencies.

Testable Constructs and Criteria for Item Selection

In studies of expertise for a variety of domains (*e.g.*, mathematics and science), two key variables that promote academic progress are principled knowledge (Alexander, 2003; Alexander, Murphy, and Woods, 1996) and reasoning with scientific or mathematical information (Cognition and Technology Group at Vanderbilt [CTGV]; Kulikowich and DeFranco, 2003; Wolfe and Goldman, 2005). One central property of reasoning is that experts are more able than novices to employ reasoning strategies when dealing with complex or novel situations they encounter within their domain (Keating, 1990). Therefore, assessing development of the cognitive processing mechanisms requires items that measure reasoning in its various forms: analogical, causal, deductive, inductive and integrative reasoning.

§ 4.2 STEM CONSTRUCTS

"Scientific thinking" is qualitatively different for physical sciences, for life sciences, and for earth sciences. The essential skill of "arguing from evidence" can rest on the ability to manipulate variables in the physical sciences (*e.g.*, the classic pendulum problem - is it the length, the weight of the bob, the angle of release, the initial "push"?) but in the life sciences it's a lot harder to apply the same kind of reasoning to data over time (*e.g.*, evolution.)

Assembling a test to measure fundamentals required for scientific reasoning across the sciences is fraught with peril because the *ability* of an individual to reason with science information is the actual *understanding* of that information. The capacity to argue from evidence (*i.e.*, the constructs that could be called "general science reasoning") is substantially correlated with the specific knowledge about that particular evidence and its context. Consequently, when the reasoning is embedded in a real context, anyone who does not know about the real context will be tripped up by the specifics inherent in that context. The richer and more interesting the problem, for example in a cross-science area like the environmental or the earth sciences, the more that information and context influences a student's capacity to follow it up.

When an assessment is severely time-constrained so that precision will be limited, the best compromise appears to be measuring some "general" effects based on a context common to all students ("ballpark measure of scientific literacy and reasoning from evidence) plus a different thing that is of personal interest to the student (second set of testlets differentiated by content rather than by global science level). The general effects may be useful for cross-sectional analyses; but the longitudinal aspects, particularly each student's growth, can be measured in that area of personal familiarity and interest. To construct such an assessment, contexts in ecology or environmental science, for example, would make it possible to frame meaningful questions about both physical and biological sciences.

§ 4.3 ITEMS

Item Selection to Address Facets

Given the importance of principled knowledge and reasoning in developing expertise, the Task Force considered item types that could be incorporated into a longitudinal, computer-adaptive testing (CAT) system for high-school students. The proposed system is to yield scores that predict declaration of college majors in science and mathematics as well as success within those programs.

Released items of the National Assessment of Educational Progress (NAEP), which can be accessed readily on the Internet, were reviewed. In science, content areas include the physical sciences (*e.g.*, physics and chemistry), earth sciences (*e.g.*, geology), and the life sciences (*e.g.*, biology). Several principles were evident in the study of the content and the problems posed within the stems of the items. These principles included: a) properties of matter; b) energy and conservation laws; and, c) ecology and population patterns. While most items were multiple-choice in format, several items were open-ended requiring students to explain a property, provide a causal explanation, or to classify information. Another group of items required students to examine information in charts, graphs, or tables and to induce or infer scientific properties. Both the content and problems posed appear to be a good initial pool of information that can be used, edited, or modified to build a system designed to analyze growth in principled knowledge and/or reasoning (see Table 2). Other released items in both science and mathematics are available for review from databases of the National Educational Longitudinal Study (NELS), Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS).

Table 2. Item types in science (NAEP released item bank, Grade 8).

Knowledge or Construct	Type of Item Content	Item Format in NAEP
Energy	Household appliances and conversion of electricity into different forms of energy.	Multiple choice.
Ecology	Food web, primary consumers	Diagram in stem. Multiple choice.
Matter	Real-world situation. What is observed when breathing on a mirror?	Multiple choice.
Data Interpretation	The relationship between pulse rate and stage of running (before, during, after).	Graphs in options. Multiple choice.
Scientific Method	How to determine which type of plant food effects plant growth the most.	Multiple choice.
Ecology with Cause and Effect	Fish in ponds.	Open ended with two parts. Part 1 requires statement of cause and effect relationship between size and frequency of fish put into a pond. Part 2 requires explaining effect due to placement of fish in the pond.
Earth and its Position in the Solar System	Comparison of shadow measurements at same time of day in Summer and in Winter.	Open ended with two parts. Part 1 requires statement of whether the measurements would be the same or different. Part 2 requires explanation for response in Part 1.

In mathematics, many of the released NAEP items for both middle- and high-school assessments focus on algebraic computation and reasoning, sometimes including geometric properties (*e.g.*, area). The algebraic computation items include factoring problems as well as representational transfer problems (*e.g.*, matching algorithmic information to coordinates or matching graphical information to equations). Many items are standard word problems requiring students to make several steps toward correct solution. Some items require reasoning with number properties (*e.g.*, odd/even, powers or primes) and this ability to reason with number properties has been indicative of emergent expertise in many mathematical specializations (*e.g.*, combinatorics and set theory; Davis, Hersh, and Marchisotto, 1995). Wainer (1990) provides an example of a hierarchical, branching strategy in the building of testlets for algebra, and several of the items reviewed in the released NAEP database may be revised to build a similar system for the longitudinal, CAT study. (See Table 3 for example item types.) To go beyond high school geometry and/or algebra II may require development and calibration of new items. Significant caution is warranted in returning to middle/early high school mathematics content when testing students who have recently been working on either advanced or alternative mathematics.

Table 3. Item types in mathematics (NAEP released item bank, Grade 8).

Knowledge or Construct	Type of Item Content	Item Format in NAEP
Number Properties in algebraic expressions	What happens to solutions of the equation, y = 4x, as x is increased by 2?	Multiple choice.
Interpreting graphical information.	Estimating the value of a point on a curve.	Graph included. Multiple choice.
Determining area.	Irregular figure with grid lines. Presented in square centimeters.	Graph included. Multiple choice.
Number series	Find the sixth number in a sequence.	Multiple choice.
Representational transfer in algebra.	Matching coordinates with equation.	Domain and range values tabulated. Multiple choice.

Representational transfer in algebra.	Matching values presented in graph with algebraic equation. Area problem.	Graph included. Multiple choice.
If-then conditional reasoning with number properties.	Even and odd number properties.	Multiple choice.

§ 4A APPLICATION TO HSLS-09

The primary goal for HSLS-09 of providing a data base for extracting an integrated picture from all sources (student, parent/family, school) of the post-secondary decision-making process, especially toward STEM trajectories, should also be useful in identifying perceived and actual barriers (*e.g.*, financial, cultural, input from the school, roles of the student and family) that new policies and/or practices might alleviate.

Success in developing this data base will require maximizing the efficiency of the test and survey instruments and focusing the available survey time on *specific issues*, probing to sufficient depth to be able to compare and contrast the perceptions of the student, the student's family, school staff and records. CAT, especially with the inaugural item levels determined from external information (whether course registration, grades, reported interest level, or in the case of 11th grade administration the score at 9th grade).

An initial decision needs to be made regarding the measurement goal, whether it is to yield a single score or a profile. In either case, a careful elaboration of the construct facets is needed, taking into account that a longitudinal study should be designed to gather information that will *augment* what is already available from external sources. With all these considerations, moving in the CAT direction is an important but challenging decision. Fortunately, learning lessons from those that have gone before should pave the way and prove valuable in achieving success.

A final criterion for the assessment is to create a data base that is not otherwise duplicated within existing record systems and that holds the potential for analysis that could not be accomplished from NAEP, SAT scores and/or information from other sources. One unique opportunity for HSLS-09 is to integrate heterogeneous information types by compiling and integrating state and school administrative records, SAT and other standardized test scores, with interest and demographic profiles and evaluation of the information available to and used by students and families in the decision process. At the same time, cross-validation of [limited] information from administrative and survey sources will also serve as a basis for determining optimal combination of information sources for the future.

It is also worth noting that there may well be information in the literature that can guide the search for relevant constructs. In particular, evaluation work, sponsored by NIGMS, has reviewed various programs funded by NIGMS, NSF, NASA, as well as foundations and other private organizations.

Science Items

The crucial question for HSLS-09 is whether or not to require a single score for science. If there is to be only one science score, then breadth of coverage is not nearly as valuable or meaningful as depth of coverage of something worth covering. Ideally, a two-part assessment could cover both general scientific reasoning and domain/context-specific reasoning, based on a context that integrates information from several domains, such as environmental science. This would allow measurement of a context-specific strength. The consequent problem is the selection bias present when some students qualify or opt for

biological science while others qualify or opt for a physical science (Wainer and Thissen, 1994). (It is possible, but difficult, to construct a design with balanced overlap to allow equating of scales across sciences. The selection bias problem and/or lack of precision from partitioning the test magnify as time is restricted and as the number of students with specific interest/knowledge decreases.)

So, if individual domain selection is not going to be possible, then the best alternative is to focus the approximately 20 minutes not on 20 contexts but on about five. For each context, the interdependence of items can be reduced by asking questions about the geology, the energy transfer, the life cycles, etc., rotating through the various science domains. *If* a second category is possible, the notion of uncertainty is nearly orthogonal, a very interesting connection for science because it addresses the concept of how "true" scientifically proven ideas are in reality.

It is important to take a broader view of persistence and attainment in STEM studies and careers. In addition to the "achievement factors," interest and motivation can come from early exposure to science, life experiences during the elementary and middle school years, perceived social value of science-based careers (especially service sector), and career expectations including financial reward. There are many policy options, for example, for increasing student opportunities to engage in meaningful science education, including out-of-school time as well as improving the science curriculum. To what extent will the data include information about students' out-of-school activities, after school programs, informal science centers and media exposure to science? It may be that the truly important linkage will turn out to be between this interest inventory plus administrative data and the science-math assessment under discussion; the two parts of the survey should not be thought about completely separately.

Mathematics Items

Mathematics assessment poses somewhat lesser, but still nontrivial difficulties. These arise principally from three sources: variable time elapsed since focused study of a topic, wide range of stages of progression through a high school mathematics curriculum, and divergence into several mathematics tracks ranging from traditional "college prep," either accelerated or not, through "business math" programs. Even in a traditional college prep program (still not elected by the majority of high school student's nationwide), the addition of a course in statistics, sometimes with alternative options such as linear algebra, can diverge from the otherwise linear track.

Assessment at earlier stages, even through middle school, can still draw on common material; and indeed, strong predictors of success in future mathematics courses have been identified. These same predictors, however, do not necessarily perform well as students advance because of the displacement of facility with old material in favor of new topics so that both successful response and speed of response can be affected.

Creating a continuum of items is difficult unless there is a basis - or a central concept - that can be used to establish relative difficulty. One viable solution might utilize an overarching topic, such as functions as relationship, as a primary basis for establishing the scale. The rationale for selecting such a topic is that this concept is fundamental to the discipline at all levels, consequently it appears prominently from first level algebra (graphing relationships) through calculus (derivative and integrals) and also appears in alternative course sequences (pricing and profit computations in business math). On course, while a single concept may be sufficient to benchmark item difficulty, the item pool needs to be more expansive.

For mathematics as for science, HSLS-09 can take advantage of administrative records (*e.g.*, most recent course content and level) to determine the initial item level (the second assessment at the end of 11th grade also use the results of the earlier assessment).

Example Items

The task force reviewed several examples of items that could be incorporated into a longitudinal CAT system. For CAT systems, multiple-choice items or one-two word (or numeric) response short answer items are often the formats selected due to ease of administration and scoring. Released items from the NAEP database that may be modified to reflect the types of principled understanding and scientific or mathematical reasoning described in this report include the following:

Science Items
Suppose that for a science project you wanted to find exactly how much the length of a shadow changes during the day. Describe both the materials and the procedures you would use to make these observations.
If you measured your shadow at noon during the summer and at noon during the winter, would the measurements be the same or would they be different? Explain your answer.
Suppose that one spring a new type of large fish was put into the pond. So many were put in that there were twice as many fish as before. By the end of the summer, what would happen to the large fish that were already in the pond? Explain why you think these new large fish would have this effect.

Mathematics Items

Which of the following equations represents the relationship between x and y shown in the table below?

a. $y = x^2 + 1$
b. $y = x + 1$
c. $y = 3x - 1$
d. $y = x^2 - 3$
e. $y = 3x^2 - 1$

X	Y
0	-1
1	2
2	5
3	8
10	29

One store, Price Pleasers, reduces the price each week of a $100 stereo by 10 percent of the original price.

Another store, Bargains Plus, reduces the price each week of the same $100 stereo by 10 percent of the previous week's price.

After 2 weeks, how will the prices of the two stores compare?

a. the price will be cheaper at Price Pleaser's.
b. the price will be the same at both stores.
c. the price will be cheaper at Bargains Plus.

REFERENCES

Alexander, P. A. (2003b). The development of expertise: The journey from acclimation to proficiency. *Educational Researcher*, 32, 10-14.

Alexander, P. A., Murphy, P. K. & Woods, B. S. (1996). Of squalls and fathoms: Navigating the seas of educational innovation. *Educational Researcher*, 25(3), 31-36, 39.

Baker, E. (2007). 2007 *Presidential Address*. American Education Research Association. 2007 GMAC Conference on Computerized Adaptive Testing, Minneapolis, MN, June 7-8.

Bethke, A., Hill, C., McLeod, L., VanDyk, P., Zhao, L., Zhou, X. & Thissen, D. (2004). *North Carolina Computerized Adaptive Testing System: 2003 Comparability Study Results*. Research Triangle Park, NC: RTI International.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika*, 37, 29-51.

Bock, R.D. (1997). The nominal categories model. In W. van der Linden & R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 33 50). N.Y.: Springer.

Bradlow, E.T. & Thomas, N. (1998). Item response theory models applied to data allowing examinee choice. *Journal of Educational and Behavioral Statistics*, 23, 236-243.

Chinn, C. A. & Brewer, W. F. (2001). Models of data: A theory of how people evaluate data. *Cognition and Instruction*, 19(3), 323-292.

Cognition and Technology Group at Vanderbilt (CTGV). Anchored instruction and its relationship to situated cognition. *Educational Researcher*, 19, 2-10.

Davis, P. J., Hersh, R. & Marchisotto, E. A. (1995). *The Mathematical Experience: Study Edition*. Boston: Birkhäuser.

Dawson, T. L. & Wilson, M. (2004). The LAAS: A computerized scoring system for small- and large-scale developmental assessments. *Educational Assessment*, 9 (3 & 4), 153-191.

Dunbar, K. N. & Fugelsang, J. A. (2005). Causal thinking in science: How scientists and students interpret the unexpected. In M. E. Gorman, R. D. Tweney, D. C. Gooding, & A. P. Kincannon (Eds.). *Scientific and Technological Thinking* (pp. 57-79). Mahwah, NJ: Lawrence Erlbaum Associates.

Edwards, M.C. & Thissen, D. (2004). Defining and Finding Optimal Designs for uMFS CATs. Paper presented at the annual meeting of the Psychometric Society, Monterey, CA, June 14-17.

Edwards, M. C. & Thissen, D. (2007). Multi-stage computerized adaptive testing with uniform item exposure. Presentation at the 2007 GMAC Conference on Computerized Adaptive Testing, Minneapolis, MN, June 7-8.

Kang, G. K. & Weiss, D. J. (2008, in press). Adaptive measurement of individual change. *Zeitschrift fur Psychologie*.

Keating, D. P. (1990). Charting pathways to the development of expertise. *Educational Psychologist*, 25, 243-267.

Koedinger, K. R. & Nathan, M. J. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *Journal of the Learning Sciences*, 13(2), 129-164.

Kulikowich, J. M. & DeFranco, T. C. (2003). Philosophy's role in characterizing the nature of educational psychology and mathematics. *Educational Psychologist*, 38 (3), 147-156.

Little, R. J. A. & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Lord, F.M. (1980) *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Luecht, R.M. & Nungester, R.J. (1998). Some practical examples of computerized adaptive sequential testing. *Journal of Educational Measurement*, 35, 229-249.

Luecht, R.M. & Nungester, R.J. (2000). Computer-adaptive sequential testing. In W.J. van der Linden & C.A.W. Glas (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 117-128). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Mislevy, R.J. & Haertel, G. (2006). Implications for evidence-centered design for educational assessment. *Educational Measurement: Issues and Practice*, 25, 58-64.

Mislevy, R.J., Steinberg, L.S. & Almond, R.G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3-67.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G. & Johnson, L. (2002). Making sense of data from complex assessment. *Applied Measurement in Education*, 15, 363-378.

Mislevy, R.J. and Wu, P.K., 1996. Missing responses and IRT ability estimation: omits, choice, time limits, and adaptive testing. *Research Report RR-96-30-ONR*, Educational Testing Service, Princeton, NJ.

Muraki, E. (1997). A generalized partial credit model. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 153-164). N.Y.: Springer.

RIttle-Johnson, B. & Star, J. R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. *Journal of Educational Psychology*, 99, 561-574.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

Samejima, F. (1997). Graded response model. In W. van der Linden & R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 85-100). N.Y.: Springer.

Scalise, K. & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "Intermediate Constraint" questions and tasks for technology platforms. *Journal of Technology, Learning and Assessment*, 4 (6) [online journal]. http://scholarship.bc.edu/jtla/vol 4/6.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, 61, 331-354.

Stocking, M.L. (1996). An alternative method for scoring adaptive tests. *Journal of Educational and Behavioral Statistics*, 21, 365-389.

Thagard, P. (2005). How to be a successful scientist. In M. E. Gorman, R. D. Tweney, D. C. Gooding, & A. P. Kincannon (Eds.). *Scientific and Technological Thinking* (pp. 159-171). Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D., Nelson, L. & Swygert, K. (2001). Item response theory applied to combinations of multiple-choice and constructed-response items—approximation methods for scale scores. In D. Thissen & H. Wainer (Eds.), *Test Scoring* (Pp. 293-341). Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L. & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement*, 26, 161 176.

Thissen, D. & Wainer, H. (Eds.) (2001) *Test Scoring*. Mahwah, NJ: Lawrence Erlbaum Associates.

Veldkamp, B.P, & van der Linden, W.J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, 67, 575-588.

Wainer, H. (2000). *Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot*. (2nd edition) Hillsdale, N. J.: Lawrence Erlbaum Associates.

Wainer, H. (2005). *Graphic Discovery: A Trout in the Milk and Other Visual Adventures*. Princeton, N.J.: Princeton University Press.

Wainer, H., Bradlow, E.T. & Wang, X. (2007). *Testlet Response Theory and its Applications*. New York: Cambridge University Press.

Wainer, H. & Keily, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, 27, 185-202.

Wainer, H. & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement*, 27, 1-14.

Wainer, H. & Thissen, D. (1994). On examinee choice in educational testing. *Review of Educational Research*, 64, 159-195.

Wang, X. Bradlow, E. T. & Wainer, H. (2004). User's Guide for SCORIGHT (version 3.0): A computer program for scoring tests built of testlets including a module for covariate analysis. *Research Report 04-49*. Princeton, NJ: Educational Testing Service.

Wang, X.B., Wainer, H. & Thissen, D. (1995). On the viability of some untestable assumptions in equating exams that allow examinee choice. *Applied Measurement in Education*, 8, 211-225.

Weiss, D. J. & Gibbons, R. D. (2007). CAT with the bifactor model. 2007 GMAC Conference on Computerized Adaptive Testing, Minneapolis, MN, June 7-8.

Wilensky, U. & Reisman, K. (2006). Thinking like a wolf, a sheep or a firefly: Learning biology through constructing and testing computational theories – An embodied modeling approach. *Cognition and Instruction*, 24(1), 171-209.

Williamson, D.M., Bauer, M., Steinbla/vol4erg, L.S., MIslevy, R.J., Behrens, J.T. & DeMark, S. (2004). Design rationale for a complex performance assessment. *International Journal of Measurement*, 4, 303-332.

Wolfe, M. B. W. & Goldman, S. R. (2005). Relations between adolescents' text processing and reasoning. *Cognition and Instruction*, 23, 467-502.

TASK FORCE

Henry Braun, *Professor, Boston College Lynch School of Education*

Kim Gattis, *Task Leader, NAEP ESSI, NAEP Assessment Development and Quality Assurance*

Jonna Kulikowich, *Professor, Pennsylvania State University, Department of Education and School Psychology and Special Education*

Robert Mislevy, *Professor, University of Maryland, Department of Measurement, Statistics and Evaluation*

Elizabeth Stage, *Director, Lawrence Hall of Science, University of California at Berkeley*

David Thissen, *Professor, University of North Carolina at Chapel Hill, Department of Psychology*

Howard Wainer, *Distinguished Research Scientist, National Board of Medical Examiners*

David J. Weiss, *Professor, University of Minnesota, Department of Psychology*

Kentaro Yamamoto, *Deputy Director, Center for Global Assessment, Educational Testing Service Research and Development Division*

Convened by the National Institute of Statistical Sciences

Nell Sedransk, *Associate Director*

For the National Center for Education Statistics

Mark Schneider, *Commissioner*

Jack Buckley, *Deputy Commissioner*

Andrew White, *Special Assistant to the Commissioner*

And for HSLS-09

Jeffrey Owings, *Associate Commissioner*