DOCUMENT RESUME

ED 380 505 TM 022 868

AUTHOR Bennett, Randy Elliot
TITLE An Electronic Infrastructure for a Future Generation of Tests. Research Report.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-94-61
PUB DATE Dec 94
NOTE 30p.; Version of a paper presented at the Annual Meeting of the International Association for Educational Assessment (Wellington, New Zealand, October 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; College Entrance Examinations; *Computer Assisted Testing; Computer Managed Instruction; *Educational Technology; Electronics; Feedback; *Futures (of Society); Licensing Examinations (Professions); *Test Construction; Test Format; Test Interpretation
IDENTIFIERS *Infrastructure

ABSTRACT

The Educational Testing Service is moving rapidly to computerize its tests for admissions to postsecondary education and occupational licensure/certification. Computerized tests offer important advantages, including immediate score reporting, the convenience of testing when the examinee wishes, and for adaptive tests, equal accuracy throughout the score scale and a shorter test with no loss in measurement proficiency. There is much more that technology can achieve, however. This paper describes an electronic infrastructure for integrating the best of traditional testing approaches with new technology. This multiorganizational infrastructure has the potential to help assessment contribute more positively to learning and decision making. It can do this by making it easier to deliver tests that employ performance tasks, include important skills not well-measured by current examinations, sample behavior frequently over a student's school career, and give instructionally useful feedback to individuals. (Contains 17 references.) (Author)

RR-94-61

RESEARCH REPORT

AN ELECTRONIC INFRASTRUCTURE FOR A FUTURE GENERATION OF TESTS

Randy Elliot Bennett

Educational Testing Service
Princeton, New Jersey
December 1994

An Electronic Infrastructure for a Future Generation of Tests

Randy Elliot Bennett

Educational Testing Service

Princeton, NJ 08541

November 29, 1994

Abstract

ETS is moving rapidly to computerize its tests for admissions to post-secondary education and occupational licensure/certification. Computerized tests offer important advantages, including immediate score reporting, the convenience of testing when the examinee wishes, and for adaptive tests, equal accuracy throughout the score scale and a shorter test with no loss in measurement precision. There is much more that technology can achieve, however. This paper describes an electronic infrastructure for integrating the best of traditional testing approaches with new technology. This multi-organizational infrastructure has the potential to help assessment contribute more positively to learning and decision making. It can do this by making it easier to deliver tests that employ performance tasks, include important skills not well-measured by current examinations, sample behavior frequently over a student's school career, and give instructionally useful feedback to individuals.

An Electronic Infrastructure for a Future Generation of Tests

Represented by the College Board's Computerized Placement Tests (for selection into developmental courses), the GRE General Test (for graduate school admissions), Praxis I (a basic skills test for prospective teachers), and the National Computerized Licensure Examination for nurses, the first generation of computer-based testing (CBT) offers several advantages over traditional paper-and-pencil measures. To start, CBT has changed the pragmatics of high-stakes test administration dramatically. Instead of taking the test on one of only a few dates per year in a large group at a temporary center, the establishment of permanent CBT centers allows examinees to test when they want, in small groups, and in more comfortable environments specifically designed for testing. Because responses are scored immediately, the examinee may see the results as soon as the test concludes.[1] In computer-adaptive implementations, tests can be considerably shorter than their paper-and-pencil counterparts, without any sacrifice in measurement precision. Additionally, because adaptive tests are dynamically built to match the examinee's skill level, substantially equal precision can be attained throughout the score scale, giving better measurement of those whose abilities fall outside the more limited target range of the conventional linear test.

These advantages are compelling to examinees and testing agencies alike. However, this first generation of CBT must be regarded as only an initial step for at least two reasons. First, this generation is limited primarily to multiple-choice and related questions, with all the potential negative consequences this implies (see Bennett, 1993, for a review of these consequences). Open-ended formats such as those that constitute the core of the performance assessment movement are minimally represented. Second, the first generation focuses on measuring traditional constructs and does not take

full advantage of the computer's potential to present stimuli and track information (e.g., through dynamic displays or recording of response latencies).

What this initial manifestation does provide is the outline of an electronic infrastructure for a future generation of tests. This paper presents one conception of this scheme. Key to this conception is combining the best of traditional approaches with new technology to form an integrated "distance" assessment system that should improve learning and decision making.

An Infrastructure for the Future

What capabilities might we expect this infrastructure to provide? Figure 1 depicts test development, test delivery, response processing, and reporting, with the order of events primarily proceeding from left to right.

Insert Figure 1 about here

Beginning on the left, today's computer-based test developers write items using a process similar to the "word-processing center" model that characterized the U.S. workplace in the early-to-middle 1980s (i.e., professional sends handwritten manuscript to center, manuscript is processed and printed copy returned, professional makes notations on copy and gives back to center, notated copy is processed and clean document returned, and so on). As in the "word-processing center" model, the CBT developer roughs out the item by entering text into the computer and drawing graphics on paper, sending both to a test production center where a CBT production specialist redraws the graphics and integrates the text to make a functioning item. The item is returned to the test developer who reviews it, communicates corrections verbally or on paper, and sends the result back to the production center.

Iterations continue until the item is as the developer and production specialist desire. In a world characterized by increasing demands for faster turn-around and multiple parallel item pools to increase test security, this process is cumbersome at best.

Tomorrow's test developer will likely use an "item-processor" to build, try out, and revise items, and package them into functional--though not necessarily operational--tests. This tool will have multi-media capabilities that permit the developer to insert not only graphics, but animation, sound, and video. Low-level screen formatting, which production specialists now spend significant time on, will be taken over by software, either as part of the item-processor interface or as post-processing. What role might CBT production specialists play? Their role might shift to (1) building and maintaining large libraries from which developers can select graphics, sound, video, and animation sequences; (2) customizing effects (that are then added to the library); and (3) making the refinements necessary to translate the developers' drafts to final product.

Such an item processor will have several interesting characteristics. First, CBT items authored in this environment will be fully functional, permitting the developer to take the item as would the examinee, or pilot test it in its draft form. Second, the tool would have generative capabilities for some item classes: Given parameters specified by the test developer, the tool would automatically rough out an item for editing. For item classes where the domain structure and the determinants of difficulty are understood, developers could create items to assess critical aspects of that structure, more confidently pitching them at desired difficulty levels.[2] Third, for open-ended items, the tool would have rubric creation aids. These aids would help the developer specify a score scale, as well as describe the features of

responses falling at each scale point. In some instances, this information would also be electronically encoded for subsequent use by automatic scoring programs. Finally, the item processor would be capable of authoring conventional, as well as computer-based, questions. This dual capability is important as conventionally delivered tasks will continue to play an important role in assessment, not only since universal access to computers will take time but because some important tasks cannot be authentically replicated in the CBT environment.

Moving to the right in Figure 1, we see that delivery will occur in "test centers." This nomenclature covers several arrangements. Students may test in large assemblages, as is common for paper-and-pencil admissions tests today; in small CBT centers; or in their classrooms. In technologically capable centers, they may take computer-based tests, conventional tests, or combinations of the two. The particular arrangement will depend on the testing program and the availability of technology.

In our vision, responses to conventionally delivered (i.e., non-computer-based) tasks, fall into two classes: Those that can be adequately converted to digital form and transmitted electronically, and those that cannot (see the top of Figure 1). The former would include all paper-and-pencil responses--whether multiple choice or open ended. Examples that might be too complex for digital capture and transmission are sculptures and live performances, though even here such solutions as digital holography, digital video, and fiber optic cable (for rapidly moving large response files), may soon prove practical. (Note that, as collections of responses, student portfolios also could be handled in this framework; whether their contents could be digitized would depend on the form those contents take.)

Responses that can be digitized are scanned, either at the test center or after being received at a central location. The digitized multiple-choice responses are then scored by a conventional program and the results reported to the examinee, one or more designated institutions, and testing organization files. Digital representations of open-ended responses (e.g., essays, mathematical proofs, diagrams) are handled differently. These are sent to computer terminals, where human judges grade them, possibly with the aid of built-in electronic tools (e.g., calculator, protractor, ruler, symbol manipulator). The judges may be at the same site as the scanner or at a site in another city, region, or country. Judges may be collocated so that they can train together and interact directly about unusual responses and changes to the rubric, or they may be distributed, communicating by electronic mail or by personal video-teleconferencing utilities.[3] Real-time moderation could occur too by having judges blindly score a common set of "anchor" responses, identifying the discrepancies, and resolving them socially or adjusting for them statistically (e.g., Braun, 1988).

Responses to tasks that cannot be digitized are handled in the usual manner. Those recorded on paper or some other non-digital medium (e.g., video tape) are physically moved from test centers to a processing location and given to human judges to adjudicate by hand. (Responses delivered live may be judged in real time, with the evaluations recorded and sent on for processing.) Last, judgments are converted to machine-readable form for use by a program that computes scores for the total test and reports the results.

As the infrastructure for computer-based testing becomes widespread, more examinees will take tests on computer and fewer through conventional means. By definition, responses to these computer-delivered tasks will be captured digitally; as a consequence, it may be possible to score some

immediately at the point of capture. Certainly, immediate scoring occurs now for responses to multiple-choice and related items (e.g., ones that require entering a numeric response). But increasingly, it will become true for more complex responses (e.g., phrases, mathematical expressions). Responses to other items will need to be transmitted electronically to a location that could, in principle, be across town or, in the case of international assessments, on the other side of the world. These responses may take the form of extended mathematical proofs, essays typed on the computer, speech captured by microphone, or diagrams constructed with a pen and tablet.

While these responses may not be fully machine scorable, some may be scorable semiautomatically. Several approaches might be taken toward such processing. In one approach, the computer would pass to human raters only those responses that it could not accurately evaluate. In those instances where scoring accuracy is related to score level, level might be used to assign responses. So, for an essay task intended to measure basic writing skill on a pass/fail scale, responses that were extremely high on automatically detectable grammatical errors (and thus almost certainly incoherent), might be failed without reading by human judges. (Those without grammatical errors might still have to be read, as the absence of such errors would not guarantee a well-written response.) In other instances, the grading program may be capable of making accurate judgments throughout the score scale. Here, it would transfer for resolution only those cases it was unsure of. This assignment strategy presumes a scoring program that can make judgments about its own performance. A second approach to semiautomatic scoring might use both machine and human graders in tandem, as when separate grades are to be awarded on different dimensions. For a persuasive essay, machine scores might be reported on such surface features as style and

grammaticality, and human scores on organization and effectiveness of argument.

Other responses will be too complex for even semiautomatic processing. These responses will be treated similarly to digitized conventional tasks. As an example, imagine a test to certify teachers of children with hearing impairments. The candidate sits down at the computer and sees on the screen a video of a person posing a question in American Sign Language. As the candidate signs a response, it is recorded by a miniature digital TV camera sitting atop the computer monitor. The response is stored, then electronically transmitted and displayed at a judge's terminal for evaluation.

When responses can be automatically scored, reporting to the examinee can be immediate. Even when responses must be transmitted electronically to another location, it may be possible to report scores unofficially by the end of the testing session if the constructed-response portions are administered first and if graders are readily available.

Regardless of when reporting occurs, it will be done electronically. In principle, it could be delivered to an examinee's electronic mail address--or simply made accessible by home computer or interactive telephony (i.e., using the telephone keypad to manipulate a remote computer). Computer-based reports will bring with them the capacity for multiple views, perhaps showing how the examinee's performance compares to established content standards or to user-defined reference groups (e.g., those with similar background characteristics, those applying to particular institutions). Finally, for those who want it, technology should make possible detailed information on the kinds of problems one was able to solve or the constellation of skills one appears to possess.

Improving Learning and Decision Making

How might this infrastructure make assessment contribute to learning and decision making? First, the infrastructure makes more practical the use of measurement methods like performance assessment. Performance assessment is common to the educational systems of many countries (Feuer & Fulton, 1994), and is finding increasing favor in the U.S. One of its defining attributes is the use of tasks that closely resemble good instructional exercises. Once reified in assessment, these tasks become positive models for teaching practice (Frederiksen & Collins, 1989).

A major impediment to using performance assessment in the U.S. has been cost. The proposed conception makes large-scale deployment more feasible by providing an integrated structure for five classes of performance task roughly arrayed by operational cost (and, not incidentally, response complexity). These are (1) computer-delivered and automatically scored, (2) computer-delivered and semiautomatically scored, (3) computer-delivered and human-scored on computer terminal, (4) conventionally delivered and human-scored on terminal, and (5) conventionally delivered and conventionally scored by human judges. Programs can opt to use whatever combination of performance tasks their educational goals require and their fiscal resources allow.

A second way this infrastructure might contribute to learning and decision making is by making it easier to measure important constructs that conventional testing programs do not now assess and which correlate only modestly with existing indicators. Including such constructs in making post-secondary admissions decisions, for example, would broaden the definition of talent and, consequently, the pool of eligible applicants.

One such construct might be "learning-to-learn," or how effectively students profit from instruction. Attempts to measure this construct have

shown promise in identifying potentially capable students who have not achieved because they have come from extremely deprived environments or have never been adequately taught (Feuerstein, 1979). Also, these measures appear to add independently to the prediction of scholastic achievement over what traditional tests provide (Campione & Brown, 1987). The general method for measuring this construct, known as "dynamic assessment," involves presenting the student with a task just above that individual's level, providing hints or other instruction, and retesting performance. For any cost-effective large-scale implementation, computer technology would be required to identify the examinee's current skill level, select appropriate tasks, control the presentation of hints, and capture the sequence of responses.

Another example might be the ability to generate alternative explanations. This skill has been judged important to success in graduate education (Powers & Enright, 1987). It also has been found to overlap only minimally with the competencies measured by existing admissions tests and to add independently over those measures to the prediction of academic performance (Frederiksen & Ward, 1978; Bennett & Rock, in press). This ability, too, could only be assessed cost-effectively on a large scale with technology.

A third way this infrastructure might contribute to learning and decision making is by making frequent behavior sampling more practical. Conventional testing programs, whether for institutional or individual accountability, typically assess performance at one point in time. For some students, this single sampling may misrepresent current capability and prospects for future accomplishment. These estimates--be they too high or too low--may encourage wrong decisions. Some of these decisions may involve learning, as in choosing between academic and technical tracks or, within

tracks, among courses of study. Better skill estimates derived from frequent behavior sampling would allow students to make more informed decisions (e.g., to pursue learning in areas best suited to them) and, consequently, increase their chances for success.

One means of implementing frequent behavior sampling is through curriculum-embedded assessment--administering tasks periodically which, because of their fit with the course syllabus, serve both institutional testing and classroom instructional purposes. To facilitate such assessment, schools could be linked to the proposed infrastructure through their own local- or wide-area networks, much as they would to any other Internet or future "Information Highway" service. Curriculum-embedded tasks might be done on computer or in paper-and-pencil and scanned before being uploaded. Once uploaded, responses that could not be automatically scored locally would be processed, not necessarily by a central authority but perhaps by teachers at other schools (using the same on-line mechanisms for rater calibration described earlier). These data would be retained by the testing program for its purposes (e.g., institutional accountability, post-secondary admissions), and could potentially become part of local, regional, national, or international databases. Schools and students would benefit from the relative unobtrusiveness of this approach, the relevance of its tasks, and the representativeness of the information it provided. All involved would gain from its faster, more cost-effective processing.

A final instance of how this infrastructure might support learning and decision making is through the type of feedback it enables. For example, tools that let developers explore alternative ways of organizing a content domain might help them better design diagnostic tests for measuring proficiencies in that domain. Second, through this infrastructure many

responses, all scores, and much other relevant information will be put into electronic form. Once in that form, these data would be available for generating individualized reports containing such elements as student profiles and digitized "work" samples illustrating standing in important areas. These reports should provide a richer picture of student accomplishment from which to gauge progress and design instruction.

Building the Infrastructure

Creating an infrastructure of this magnitude is likely to require a multi-organizational effort. Rather than being built anew, it would use existing (or future) computer networks created for more general purposes. In all probability, this assessment infrastructure would be only part of a larger, integrated series of services. The other services might include test registration, information, and preparation; career and academic guidance; application to post-secondary education; and instruction. These additional services would be accessible not only from school, but from home.

Whereas this next-generation infrastructure is a long way off, rudimentary portions already exist, some in prototype and some in operational form. Looking first at the part of Figure 1 that deals with computer-based tests, the most substantial extant component is the ETS/Sylvan Learning Systems network, which now comprises over 250 operational centers in the U.S. and abroad, and which should double by the 1996 academic year. Item pools and software updates routinely flow electronically from ETS to these centers, while responses to multiple-choice questions, to simple constructed-response items, and essays written on computer pass back to ETS on a daily basis. From ETS, scores are reported electronically to some test sponsors. In the case of the National Council of State Boards of Nursing, reporting occurs within 48 hours of the examinee's test administration.

This network does not yet have the ability to score performance tasks automatically beyond those involving literal matches or simple equivalencies (e.g., fractions to decimals). The first use of automatic scoring will be January 1995, with the experimental introduction of the "expression" response type.[4] This response type encompasses the class of test questions whose answer is a single mathematical expression. Examinees enter the expression by using a mouse to click on a series of symbols and numbers. In general, there will be an infinite number of mathematical paraphrases for the correct response to any member of this item class. Based on symbol manipulation algorithms found in such off-the-shelf software as Mathematica, the scoring program evaluates the examinee's response in real time to determine if it is algebraically equivalent to a test-developer key.

More complex interactive performance tasks are in the prototype stage. For example, Bennett and colleagues (Bennett & Rock, in press) have created a computer version of the Generating Explanations task, which was administered experimentally with the computer-adaptive GRE General Test through the ETS/Sylvan network. Kaplan has attempted to automatically score the responses using a program for understanding phrases and single sentences (Kaplan, 1992; Kaplan & Bennett, 1994). Results suggested the program was not sufficiently accurate to use in a fully automatic manner because of the divergent nature of Generating Explanations responses (i.e., many plausible explanations exist for each situation). Building on an idea by Kaplan and Bennett (1994), Kud, Krupka, and Rau (1994) have made a semiautomatic procedure. Their program is of particular interest because it generates for every response it can score an index indicating the confidence it has in the evaluation assigned. Thus, responses that cannot be scored, as well as those about which the program has low confidence, can be routed to human judges for interactive resolution. The

program then automatically uses the judge's entry to update its scoring rules, permitting it to process other instances of the same response without intervention.

Also in prototype form is the Advanced Placement Computer Science (APCS) Practice System (Bennett & Wadkins, 1994). This system currently is run (apart from the ETS network) by students taking a college-level computing course in high school. The APCS system contains over 50 programming tasks covering one segment of the course curriculum. Students can solve problems, execute their solutions with system-generated test data, and, for some problems, receive automatically produced partial-credit scores and diagnostic comments. The system logs various information as the student solves the problem--compilations, executions, scores, diagnostic comments, and source code--providing a trace of the solution process. It is possible that automatic analysis of this trace might identify for intervention students who utilize unsystematic (and ineffective) solution strategies. Also, because the system is meant to be used over time, scores from multiple behavior samples could possibly supplement the culminating examination now employed by post-secondary institutions to award advanced placement and/or course credit. Finally, it is conceivable that dynamic measures might be derived from how effectively the student uses feedback from program compilation, execution, and automatic grading.

For test development purposes, the ETS network does support some rudimentary computer-based tools, though as suggested these follow the "word-processing center" model; that is, they do not allow real-time creation or revision by the test developer. Katz and Zuckerman (1994) have developed FRADSS (Free-Response Authoring, Delivery, and Scoring System), a prototype item processor that permits test developers to construct computer-based items

from "objects." Each object brings with it capabilities that enable the item to behave in certain ways (e.g., present an animation), or the examinee to act upon it (e.g., draw a line, shade a portion of a figure, move figures). The developer can create items from various combinations of objects, interact with them as would the examinee, revise them in real time, assemble a test, and deliver it in pilot form. At present, this tool can be used only to prototype and pilot-test items; once selected for operational use, items must be reimplemented in the ETS OSA (Open Systems Architecture) software environment. A version of the tool that should interact with the ETS environment is now being developed.

Singley and Bennett (in press) are building a domain-specific tool for designing mathematical reasoning tests comprised of constructed-response word problems. The distinguishing features of this tool are that it provides an organization of the domain, assists the test developer in defining a partial-credit scoring rubric for each item, and generates from the rubric much of the information needed for automatically scoring responses. The domain organization is based on a structural analysis of problems. As such, it gives an overview of the problem space, allowing the developer to design tests that cover whatever portions of that space are considered important for a given assessment purpose. By designing tests in this way, the developer builds in the potential for feedback describing the examinee's facility with specified segments of the domain.

Turning to the portion of Figure 1 that represents conventionally delivered tests, many organizations have well-honed processes for scoring paper-and-pencil, as well as more complex, performance tasks. These processes center around gatherings of human judges to develop rubrics and evaluate responses. National Computer Systems (NCS) has automated part of the process

and has used it operationally to score several million constructed-responses to the U.S. National Assessment of Educational Progress (NAEP), as well as those from other programs. Paper-and-pencil responses (e.g., to essay prompts, mathematical problems) are shipped to a central facility, where they are scanned, digitized, and uploaded to a wide-area network. The digitized representations are given to human judges at terminals, who are grouped according to discipline. Because the images are digitized, the judge can enlarge any portion of the response. The system can assign every nth response to multiple judges as a means of checking rater agreement, and the table leader can view the results in real time. This real-time analysis allows the table leader to stop the scoring to clarify elements of the rubric with individuals, or with the group, as needed. Because each group is composed of about a dozen individuals housed in a private room, the collegial interaction is much the same as it is in a paper-based scoring session, with frequent exchange about such things as unusual responses and the rubric modifications needed to accommodate them. Results of using the system with NAEP responses suggest that, while rater reliability levels are comparable to the conventional method, more responses are scored per unit time and much less labor is needed to manage the movement and storage of responses (J. Goodison, personal communication, September 9, 1994).

With respect to the reporting end of Figure 1, the Graduate Management Admissions Test is perhaps the first major testing program to include performance samples as part of its score reports ("New GMAT," 1994). In addition to paper-and-pencil multiple-choice verbal and quantitative sections, the GMAT now includes two handwritten analytical essay tasks. The examinee's response to each is digitized and a copy appended to the numerical test

results, allowing admissions committees to evaluate the quality of the candidate's reasoning and writing skills directly.

Conclusion

This paper presented one conception of a multi-organizational infrastructure for a future generation of tests in which conventional and new technological capabilities are combined to form an integrated "distance" assessment system. The system is a distance one in that examinees, graders, test developers, CBT production specialists, response-processing programs, and score recipients might reside at different locations but be linked electronically.

This infrastructure should help assessment contribute to learning and decision making. It should do this by making more feasible tests that (1) employ performance tasks modeling good instructional practice, (2) include important skills not well-measured in current examinations, thus broadening the criteria upon which assessment decisions are made, (3) sample behavior frequently, providing proficiency estimates that help students better plan their schooling, and (4) give feedback to facilitate individual growth.

This electronic network might be part of a larger arrangement delivering additional, but integrated, educational services--test registration, information, and preparation; career and academic guidance; application to post-secondary education; and instruction. Making it easier for examinees to take tests, helping them perform in a manner that accurately reflects their capabilities, giving them the guidance needed to make good decisions, and making the application process as painless as possible can only further improve assessment.

Significant portions of the proposed assessment infrastructure exist, but in a fragmented and, often, experimental state. Our major challenge will

be in creating computer routines to make a wider range of tasks automatically scorable, moving current prototype capabilities to production, and integrating the various infrastructure components into a coherent whole that ultimately helps institutions and individuals make better educational choices.

References

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. J. Mislevy, and I. Bejar (Eds.), Test theory for a new generation of tests (pp. 323-357). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.). Construction vs. choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 1-27). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bennett, R. E., & Rock, D. A. (in press). Generalizability, validity, and examinee perceptions of a computer-delivered Formulating-Hypotheses test. Journal of Educational Measurement.

Bennett, R. E., & Wadkins, J. R. J. (1994). The Advanced Placement Computer Science Practice System: Development of an instructional assessment model (RM-94-20). Princeton, NJ: Educational Testing Service.

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. Journal of Educational Statistics, 13, 1-18.

Campione, J. C., & Brown, A. L. (1987). Linking dynamic assessment with school achievement. In C. S. Lidz (Ed.), Dynamic assessment: An interactional approach to evaluating learning potential (pp. 82-115). New York: The Guilford Press.

Feuer, M. J., & Fulton, K. (1994). Educational testing abroad and lessons for the United States. Educational Measurement: Issues and Practice, 13(3), 31- 39.

Frederiksen, N., & Ward, W. C. (1978). Measures for the study of creativity in scientific problem-solving. Applied Psychological Measurement, 2(1), 1-24.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. Educational Researcher, 18(9), 27-32.

Feuerstein, R. (1979). The dynamic assessment of retarded performers: The Learning Potential Assessment Device, theory instruments, and techniques. Baltimore: University Park Press.

Kaplan, R. M. (1992). Using a trainable pattern-directed computer program to score natural language item responses (RR-91-31). Princeton, NJ: Educational Testing Service.

Kaplan, R. M., & Bennett, R. E. (1994). Using the Free-Response Scoring Tool to automatically score the Formulating-Hypotheses item. Princeton, NJ: Educational Testing Service.

Katz, I. R. & Zuckerman, D. I. (1994). A software tool for rapidly prototyping new forms of computer-based assessment. Unpublished manuscript. Princeton: Educational Testing Service.

Kud, J. M., Krupka, G. R., & Rau, L. F. (1994). Methods for categorizing short answer responses. In R. M. Kaplan and J. C. Burstein (Eds.), Educational Testing Service Conference on Natural Language Processing Techniques and Technology in Assessment and Education. Princeton, NJ: Educational Testing Service.

"New GMAT adds essay component.' (1994, November). ETS Access. pp. 1-2.

Powers, D. E., & Enright, M. K. (1987). Analytical reasoning skills in graduate study: Perceptions of faculty in six fields. Journal of Higher Education, 58, 658-682.

Singley, K. M., & Bennett, R. E. (in press). Automatic scoring of mathematical expressions using symbolic computation (RR-xx-xx). Princeton, NJ: Educational Testing Service.

Author Notes

This paper is adapted from a presentation at the annual meeting of the International Association for Educational Assessment, Wellington, New Zealand, October 1994.

Appreciation is expressed to David Kuntz, Kevin Singley, Len Swanson, and Bill Ward for their helpful reviews of an earlier version of this paper.

Footnotes

1. Scores presented immediately after testing are unofficial. Some programs have elected to forego this presentation and instead promptly deliver official score reports by mail.

2. See Bejar (1993) for more on the concept of item generation.

3. By outfitting one's personal computer with a miniature TV camera, a microphone, and the appropriate software, video-teleconferencing among several parties simultaneously can already be conducted over the Internet.

4. Creators of this response type included, among others, Kevin Singley, Dave Bostain, Daryl Ezzo, Jutta Levin, Mary Morley, Alex Vasilev, and Randy Bennett.

Figure Caption

1. An infrastructure for a future generation of tests. Note. Conv. MC = response to a conventional multiple-choice task; CBT-MC = response to a computer-based-test multiple-choice task.

Development Delivery Processing Reporting