COMPUTERIZED DIAGNOSTIC TESTING: PROBLEMS AND POSSIBILITIES

David L. McArthur

CSE Report No. 255
1985

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

APPLIED STUDIES IN COMPUTERIZED DIAGNOSTIC TESTING: IMPLICATIONS FOR PRACTICE

ABSTRACT

The use of computers to build diagnostic inferences is explored in two contexts. In computerized monitoring of liquid oxygen systems for the space shuttle, diagnoses are exact because they can be derived within a world which is closed. In computerized classroom testing of reading comprehension, programs deliver a constrained form of adaptive testing and error performance summary. However, the world is open: diagnostic inferences cannot be made with precision, and additional practical factors play an important role in delimiting the usefulness of such a system. Problems of uncertainty, negation, and nondeterministic prediction are also discussed.

Introduction

Because of modern computer hardware and software, an intelligent system for diagnostic testing which incorporates the advantages of computerized management with the latest theoretical developments in diagnostic test strategy is no longer locked in the world of science fiction. In theory, a small computer could manage an individualized adaptive testing session, drawing on a bank of diagnostically relevant test items, making real-time decisions about competing diagnostic hypotheses based on the incoming stream of responses. In theory, even if premised on a rough set of diagnostic indicators, such a system ought to generate a functional summary of the performance of an examinee. Because the task of diagnosis in its most elementary form is simply one of identifying consistent patterns of examinee behavior, it seems an ideal task for the computer.

In reality, of course, neither does the naive view of the diagnostic process portrayed above hold true for a moment, nor does blind application of high-technology computer programming circumvent an array of decisions about the nature of performance and its context, the structure of performance testing, and a virtual guarantee of multiple uncertainties in interpretation. Important problems arise in programming a pattern-diagnostic inferencer to diagnose performance as it occurs, in operating that program, and in deriving meaningful diagnostics from its outcomes. Finally, even in the best of circumstances, improvements in the computability of diagnostic testing hinge on developments in computer software and diagnostic theory which have yet to occur.

Lest the reader feel that this viewpoint is unduly pessimistic, we note that computerized diagnostic testing is functioning at this moment in fields as diverse as reading comprehension and the launching of space vehicles. Their salient features and extensions to educational diagnostic testing using computers, are the subject of this paper. Because of its success, the space shuttle diagnostic system can be used to illustrate critically important conceptual underpinnings of the diagnostic process, which are generally lacking from diagnostic strategies in education and psychology.

The earliest attempt at what would now be called computerized diagnostics was generated by a "teaching-learning machine" designed by Pressey (1926). A rachet-driven device, not unlike a manual typewriter, presented selected test items in a viewing window; responses were made on a specialized keyboard and scored mechanically. The process was envisioned as labor-saving, to "leave the teacher more free for her most important work, for developing in her pupils fine enthusiasms, clear thinking, and high ideals" (p.376). More recent work in computerized diagnostic testing in educational settings has been discussed in Bejar (1984), McArthur and Cabello (1985), McArthur and Choppin (1984), Mitchell (1982), and Schwartz (1984). In reference to computerized psychological testing, Roid (1985) presents an extended overall analysis, though sketchy on the issue of diagnosis. To summarize, all these writers agree that the potential of computers applied to the particular tasks of administering scoring and supplying the bases for test interpretation looks genuinely good. Indeed, a large amount of computer code to accomplish computer-managed testing is

included in Schwartz's book, and several commercial test publishers have begun marketing aggressively in this area.

One reason for optimism is the power of the latest generation of small computers. Computer hardware was formerly a major bottleneck in implementing computerized diagnosis. Not long ago, few machines were capable of handling the job without severe restrictions on speed, memory, storage, and ancillary capabilities. In less than a decade, computer technology has leapt forward in ways which now allow extraordinarily complex logical and mathematical operations to be implemented very rapidly. Most restrictions that used to apply are gone, since reliable hardware can now include not only keyboard and video display, but also voice synthesizer, voice recognition device, and real-time graphics. Highly veridical problem simulations are now possible. Alternatively, if the testing is only a matter of presenting text to an examinee and waiting for a keystroke response, then modern lap-top computers suffice nicely. In sum, hardware no longer poses a significant barrier to the development of diagnostic tools.

The task of computerized diagnosis is much more demanding on computer software. Both logical and mathematical operations must work in an environment of real-time (respond now to this test item) and periodic (save the examinee's response pattern in long-term memory) operations. Fortunately, a number of programming languages are equipped to handle these composite requirements. An important software problem which is more difficult to solve is the handling of exception conditions. Exceptions occur when the program encounters some action or data which it is not

prepared to handle. While some languages return a nil, a default, or an explicit "don't know," others respond to that event with total program failure. When exception handling is added to the requirements noted above, no single programming language emerges as the perfect software vehicle for computerized testing. Ideally, a real-time oriented language for programming of computerized diagnosis would include extensive facilities for error-trapping as well as both symbolic manipulation and arithmetic computation. Even advanced languages like Modula-2, C, and LISP, and CAI production systems like PILOT, incorporate only some solutions to these issues, so the final choice awaits further developments in software technology.

A Closed-World Diagnostic Inferencer

The space shuttle launch monitoring system described by Scarl, Jamieson and Delaune (1985) serves as an excellent model for computerized diagnostics, on the one hand because it is highly effective and on the other because the world in which it works is well formulated. Space shuttles are launched under extraordinarily tight controls, with thousands of critical indicators being monitored and evaluated continuously by computer. Recently, the monitoring of liquid oxygen activity (valves, pipes, tanks, flow rates, pressures and the like) has been accomplished by a computerized expert diagnostic system, operating as an intelligent watchdog, with the capability of quickly isolating and interpreting any error anywhere within its purview. Its diagnostic strategy is strongly predicated on the notion that a truly simultaneous occurence of independent errors is highly unlikely. Far more probable is a failure in some

component which has consequences felt immediately or soon thereafter in several places further along the chain.

What happens when the liquid oxygen expert "discovers" that one or several sensors are reporting values out of normal range is what makes it an excellent expert to study from the point of view of diagnosis. In the simplest case, the system receives indication of a single-point error: a single sensor registers abnormally high or low. The expert "knows" enough to assess the degree of criticality of that component, and to "understand" whether failure at that point in the complex array of liquid oxygen circuitry should have consequences felt downstream by other sensors. If all downstream indicators are reporting clear the most likely explanation of this single erroneous indication is sensor failure.

If, on the other hand, a cluster of errors is suddenly reported together, the expert evaluates the root cause of such multi-point failure in two ways. The first is a method of set intersections, using assumptions about the state in which matters would have to be in order to produce those values being received at this time. The second is a method based on simultaneous hypothesis testing, using the logic of a propagating error tree which is tested in increasing depth until a point source of the error is isolated.

The liquid oxygen diagnostic system operates in a closed world. Its sensors cover the entire domain, and errors within that domain are registered unambiguously. As long as the system's programmers have properly placed each sensor and have accounted for any unique operating characteristics or "quirks," no error of any consequence whatsoever will go

undetected. For such a system, the world beyond its sensors need not be considered because every plausible diagnostic possibility has already been included.

Assumptions for the Closed World

The requirement for a closed-world diagnostic system is a test domain in which all possible faults may be enumerated discretely, and in which each single source of data may be pegged, in advance, as to its range of reporting values. The introduction of a single fault not contained in the list of known faults, or of a single datum of unknown character, exceeds the closed world and destroys its advantage. A key part of the advantage of a completely closed world is that all of the operating characteristics of that world can be known exactly. They need not be explicated in entirely, but by virtue of their availability, the closed-world inferencer has the resources to evaluate any plausible permutation of events.

Suppose we are interested in diagnosing faults in a contained domain like the liquid oxygen system with the constraint that we do not yet know exact tolerances for many of the sensors. We could proceed by discarding the evidence shown by those sensors altogether and instead use only those pieces of evidence about which we have advance knowledge as to its shape. We could allow the shuttle to be launched under a series of controlled trials regardless of the data until we amass a repertoire of interrated cause-and-effect relationships between sensor reports and final outcome, using that experience to build a library of allowable values. We could attempt to corroborate the multiple data from sensors from the present

shuttle with salient aspects of previous experience, then use such experience as a selective and conditional guide to completing the present task. Multiple avenues could be productively explored given enough resources, such that a suitable diagnostic evaluation eventually could be made of liquid oxygen system activity. Obviously, however, the operational advantage lies with the system which need not engage in strictly exploratory behaviors before being able to form diagnostics conclusions.

Two additional considerations must also be made: the first concerns the nonintermittency of signals while the second concerns the granularity of the data received. Nonintermittency is a strong assumption within the liquid oxygen diagnostic system. While each sensor is capable of generating a continuous stream of data, sensing of the status of any given sensor occurs at discrete intervals. Any sensed value is expected to be regular. That is, stable readings are seen as far more likely, from the point of view of diagnostic interpretation, than are wild fluctuations within short intervals. Indeed, intermittent fluctuations are more readily interpretable as sensor errors and noise than as diagnostically relevant indicators.

The second additional consideration is that the data received from the various sensors are at the functional granularity demanded by the diagnostic inference process. This means, on one hand, that no aggregation of incoming values need be made prior to using them diagnostically, and, on the other, that no step in the diagnostic process will require finer shades of data than are being delivered. The granularity of data, in this instance, is optimal.

Diagnostic Inference in an Open World

While there are several approaches which have been taken to building computerized diagnostic testing in the domain of reading comprehension, few of the assumptions used in closed world diagnostics carry over. Reading comprehension is a domain for which few theorists, if any, have attempted to formalize all of the likely diagnostic indicators of erroneous performance. The diagnostic inference process in reading comprehension, even as practiced by professionals, represents more than simple rules of procedure and logical chaining of consequences. Characteristically the process reflects an accumulation of overlapping evidence, plus both common sense and, for lack of a better descriptor, professional acumen. Neither of the latter factors are especially amenable to computerization. Nonetheless, computer programs now exist which are capable of adaptively presenting a limited scope of reading test items and deriving from the response pattern a composite error summary.

In a domain such as reading comprehension, the scope of a student's misunderstanding can be quite large, so large as to make it exceedingly difficult to predict all possible errors. The likelihood of a single-point error is slim, since so few errors in reading are unitary. Most often an examinee will demonstrate multiple errors, yet isolation of a single cause of a multi-point error cannot use a system of tracing error propogation because no theory of reading yet includes one. It is rather unlikely that the data from such testing can be construed as always nonintermittent, and even more unlikely that the raw responses are at an optimal level of granularity. For computerized diagnostic testing of reading comprehension the benefits of closed world assumptions do not hold. The relatively

simple rules which allow noisy data to be cast out cannot be applied. Even in the most carefully prepared of the present systems for appraising reading comprehension skills, what constitutes a diagnostically useful pattern of erroneous responses is not completely resolved.

Our experience with a dedicated test administration/feedback system and a test of reading comprehension skills specifically designed around diagnostic principles demonstrates some positive outcomes despite the concerns portrayed above. One hundred and sixteen upper primary pupils were given a pair of brief computer-managed tests, one a test of pronoun usage and the other a reading test involving short essays. Items were calibrated by difficulty and item distractors were keyed as to the type of error each reflected. Movement from items of moderate difficulty to items of greater or lesser difficulty were controlled by real-time appraisal of examinee performance. This movement up or down was significantly related to examinee grade level and reading skill. An examinee's movement up or down in one test was generally corroborated by the same movement up or down in the other test. The pronoun test, which represents one of the more closed worlds in reading skills, showed a fair degree of performance consistency within examinees, and a balanced and logical distribution of error types by skill level across examinees. The comprehension test, representing a more open-world domain, showed a somewhat less logical error pattern overall: students who evidently had the capacity to properly answer items at a middle range of item difficulty frequently stumbled on simple errors of literal comprehension when answering more difficult items.

Practical Requirements

Our testing with a prototype diagnostic inferencer in classroom settings suggests that a number of the troubles noted in reference to diagnostic strategy in the open world can be favorably resolved. Heavy emphasis must be given to designing a test which adequately covers the full scope of a given topic, and does so with items which possess good psychometric qualities and well-formed error categories. Based on technical considerations, it should be pointed out that there are other requirements that dictate the nature of a test which would be suitable for computer-managed diagnostic testing in education or psychology.

The first requirement is that the test match the operations of the computer: its text must fit within an available screen window, its tasks (type your name, hit this response key) must be unambiguous to the user, its options (strike this key to go forward or that key to go backward) must be exceedingly clear. The default instructions for taking a paper-and-pencil test, so thoroughly ingrained in most students by habit alone, are not automatically transferred by students to computer testing. For students who falter as they go to type their name at the keyboard before the test begins, frustration already mounts.

Second, the test as a whole must be user-safe. That is, the software must be "fire-walled." At the opening requirement that the student type their name at the keyboard, there are dozens of possible variations which must be distinguished by the computer from an incomplete or erroneous attempt; the software must only move forward when the student is ready. No matter what logical or illogical key sequence is pressed, the software must

be able to separate a legitimate response from careless keystrokes or lightly malicious attempts to "fool the system." Few students will try the latter, but nonetheless a computerized test which succumbs to one pupil's errant behavior will stand little chance of surviving the remainder of the allotted time period, simply because students find killing the system a great deal more interesting than completing the test.

Third, the test must be intrinsically interesting, separate from the novelty of its appearance on a videoscreen. Because the experience and excitement of videogames is almost universal among school children even at the lowest grades, expectations of what a computer will do are now jaundiced. Repeated presentation of chunks of text on screen, with a single keystroke as the sole behavior required of the student, is deadly. Some students may find themselves striking a response key simply to make the screen do something -- anything; in that instance, all of the ordinary concerns about random and partially random responding to test items are exacerbated. The results of computerized testing and any ensuing diagnostic interpretations become uncertain at best.

Models for Handling Uncertainty

A diagnostic inferencer which works in anything but a completely closed-world environment runs headlong into issues of uncertainty. The variety of ways in which uncertainty can be handled statistically and probabilistically suggests that no single solution suffices. Pearl (1984) describes detailed logical and mathematical approaches to the task using an approach which stems from a Bayesian tradition. Prade (1985) models imprecision and uncertainty with a deductive system modeled on fuzzy-set

logic. Reggia, Perricone, Nau, and Peng (1985) delineate an abductive inference process in which plausible causal associations are derived sequentially by testing symbolic conditional probabilities through set theory.

In the context of diagnostic testing, the ways in which uncertainty is managed are crucial to the diagnostic outcome. An example of this is negation, deciding when a particular diagnostic hypothesis is no longer viable. Cohen's (1984) multiple definitions of rule endorsement and negation form a case in point. The weakest interpretation of negation (called "ostrich") allows a hypothesis to be negated if positive evidence in support of the hypothesis does not currently appear in the data. The strongest interpretation of negation (called "hard-not)" requires hard evidence in support of the negation of a hypothesis or in support of the hypothesis' opposite be present in the data. A closed-world assumption requires that evidence for negation of the hypothesis be present or that a proof offered in support of the hypothesis fails.

Almost all of the operations undertaken in a diagnostic test in education are subject to negation at one point or another.
The task of diagnostic testing can be interpreted as an exploration of competing hypotheses and a weeding out of those hypotheses which are not receiving support. Because of the uncertainties inherent in test responses, the removal of a plausible hypothesis from the set of hypotheses under study is seldom matched by strong evidence. Most frequently, one has to make use of weak negative information to place that hypothesis on the back burner, then rely on good luck to isolate diagnostically important

information from the hypotheses which remain. The problem is one of logistics as well as mathematics: how can one optimize the selection of avenues to explore without predicting that some paths are likely to be less fruitful than others? In diagnostic testing in an open-world environment, these predictions are very difficult to make.

In a parallel domain, a recent contribution to the field of artificial intelligence sets out a series of transformation methodologies to determine sequence-generating rules (Dietterich & Michalski, 1985). The problem, a direct analogue of real-time diagnostic testing, is one of predicting future behavior by looking back, in varying degrees of depth, at the evidence so far -- that is, estimating what constitutes a meaningful summary pattern and expectation for the next behavior in sequence based on some or all of what has gone before. The simplest version of this problem occurs when the next behavior (or object, or response) is one which is totally predetermined by every attribute associated with all past objects. All of the attributes in the preceding string of evidence can be used to form a perfect prediction of what will come next; the methodological problem reduces to counting the total number of distinct attributes.

In a nondeterministic prediction problem, the occurence of the next piece of evidence may or may not entirely fit the string of evidence collected to date. Certain subsets of attributes may play more significant roles in determining the next evidence than others. The goal is to find plausible and parsimonious descriptors of key patterns underlying the evidence. This is a close equivalent to the kind of diagnostic process seen in the non-closed-world systems discussed earlier. The methodological

problems are significant, for the solution requires understanding which attributes of the evidence collected to date actually contribute meaningfully to behaviors, and which attributes are misleading, irrelevant, and/or simply reflect random values.

Dietterich and Michalski (1985) suggest that the solution to non-deterministic prediction might rely on any of three approaches to variable-valued logic calculus: disjunctive-normal modeling, decomposition modeling, or periodic modeling.* When objects or events in the opening sequence make use of a small set of finite-valued attributes, experimental evidence suggests that objects or events later in the sequence can be predicted well by any of the three models, looking backward to various depths at the preceding evidence. Given a stream of data which defies categorization, all of these models will either labor for extended periods of time without producing useful results, or "discover" one rule or another which fits the data badly. Unfortunately, current implementations of all three methods suffer from the weakness that they do not attempt to evaluate

---

*A disjunctive-normal model builds upon "the fewest number of conjunctive terms that covers all of the positive examples and none of the negative examples" (p.219). An iterative process generates an increasing number of maximally-general expressions until no positive example remains which is not already covered. The decomposition model iterates trial versions of generalizations of attributes among pieces of evidence. Its intermediate results are then tested against the negative evidence found in the data, and it concludes only when the decomposition succeeds in excluding all negative evidence. The periodic model expands on this latter approach, testing conjuncts of positive evidence and comparing the degrees of overlap between attribute selectors until some functional minimum of overlap is reached, and at the same time no negative evidence remains included by the hypotheses.

the best solutions first, they have no way to assess the plausibility of their results, and they cannot at present form composite models.

Summary

Computerization of the diagnostic testing process in education is a challenge of multiple dimensions, including both operational aspects and philosophical underpinnings. Indeed, the question has been raised as to whether computer software can adequately represent the subtle but important common-sense elements which come into focus when the domain of interest is not within a closed world (Bobrow & Hayes, 1985). As the preceding analysis has shown, the closed world provides a substantially cleaner environment within which to perform diagnostic inference. In the case of educational diagnosis, most domains tend to be relatively open-ended and thus no comparable clarity can be found.

If the test materials for computerized administration can be designed within tightly controlled parameters, and if the diagnostic strategy can be strongly tied to theory about performance errors within the topic domain, then many of the ambiguities of diagnostic inference will be closer to resolution. The algorithm that is used to select the next item in sequence is also critical: along with item calibrations, a selection algorithm could use diagnostically probative items, items which are particularly suited to explore the examinee's misunderstanding of a given concept within the text. Ideally, too, the pattern of erroneous performance of an individual respondent, instant by instant, could be analyzed in the context of similar patterns generated in previous testing sessions.

Yet to be solved is the problem of diagnostic precision. Inherent in most educational topics are unstudied assumptions about the ways in which erroneous performance manifests itself. Psychometric evidence makes clear that patterns of errors in multiple-choice test items occur in complex ICC distributions which are neither directly interpretable from theory, nor completely orthogonal to other traits of the test or the respondent. Thus, at present, even with the best of computerized testing, the veracity of diagnostic outcomes from computerized testing must be closely scrutinized.

The computer has proved itself valuable in managing more traditional varieties of educational test administration and scoring. Properly programmed, the computer can become an unparalleled asset in the context of diagnostic testing, if certain limits are observed. Taken collectively, the sheer number of limits both of a philosophical nature and in reference to actual testing practice strongly suggest that the computer's role will be supplementary to the educational diagnostic specialist. Breakthroughs, however, could occur as soon as computer software moves into its next generation of power, and as soon as educational theorists are able to build detailed models of misunderstanding.

REFERENCES

Bejar, I. (1984) Educational diagnostic assessment. Journal of Educational Measurement, 21, 175-189.

Bobrow, D. G., & Hayes, P. J. (1985) Artificial intelligence -- where are we? Artificial Intelligence, 25, 375-415.

Cohen, P. R. (1984) Heuristic reasoning about uncertainty: an artificial intelligence approach. Boston: Pitman Advanced Publishing Program.

Dietterich, T. G., & Michalski, R. S. (1985) Discovering patterns in sequences of events. Artificial Intelligence, 25, 187-232.

McArthur, D. L., & Cabello, B. (1985) Diagnostic testing: Field trial results. NIE Final Report, Grant No. NIE-G-83-0001, P1.

McArthur, D. L., & Choppin, B. H. (1984) Computerized diagnostic testing. Journal of Educational Measurement, 21, 391-397.

Mitchell, A. C. (1982) Using microcomputers to help teachers to develop their assessment procedures: A development project report. Programmed Learning and Educational Technology, 19, 251-256.

Pearl, J. (1984) Heuristics: Intelligent search strategies for computer problem solving. Reading, MA: Addison Wesley.

Prade, H. (1985) A computation approach to approximate and plausible reasoning with applications to expert systems. IEEE Transactions on Pattern Analysis and Machine Intelligence, 7, 260-283.

Pressey, S. L. (1926) A simple apparatus which gives tests and scores -- and teaches. School and Society, 23, 373-376.

Reggia, J. A., Perricone, B. T., Nau, D. S., & Peng, Y (1985) Answer justification in diagnostic expert systems - Part I: Abductive inference and its justification; Part II: Supporting plausible justifications. IEEE Transactions on Biomedical Engineering, 32, 263-272.

Roid, G. H. (1985) Computer technology in testing. In B. S. Blake & J. C. Witt (Eds.) The future of testing: The second Buros-Nebraska symposium on measurement. Hillsdale, NJ: Erlbaum.

Scarl, E., Jamieson, J. R., & Delaune, C. I. (1985) Process monitoring and fault location at the Kennedy Space Center, SIGART Newsletter, 93, 38-44.

Schwartz, S. (1984) Measuring reading competence: A theoretical prescriptive approach. New York: Plenum.