INTERACTIVE DIAGNOSTIC TESTING:
FIELD TRIAL RESULTS

David L. McArthur
Beverly Cabello

CSE Report No. 254
1985

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

ABSTRACT

DIAGNOSTIC TESTING: FIELD TRIAL RESULTS

David L. McArthur and Beverly Cabello

UCLA Center for the Study of Evaluation

A diagnostic testing system managed by microcomputer was evaluated in actual use at the upper primary level. Two tests specifically designed to yield diagnostic indicators of erroneous performance were utilized, one a test of pronoun usage, the other a test of reading comprehension. The results are interpreted from the standpoint of the examinees, of the tests, and of the computer software. Lessons about the viability of test management software in real-time use, about the patterning of erroneous responses, and about the efficiency of diagnostic testing by computer are discussed.

INTRODUCTION

There are few problems as widespread among primary school pupils as the inability to comprehend written materials. However, in the realm of diagnostic testing of reading comprehension, progress has been limited. Generally a student is asked to take a test from beginning to end, and only upon completion does diagnostically useful information about that student's abilities begin to emerge. Frequently the interpretation is based solely on percentage scores derived by relatively elementary algorithms concerning tallies of right and wrong responding.

If the student encounter with the test could be guided by an "actively involved onlooker," the test itself could proceed selectively, and the student encouraged to tackle questions which are increasingly "well-suited" in the sense that they become increasingly closer estimates about the student's optimal level of functioning. Each estimate would be supported by one or more diagnostic hypotheses, with suitable confidence levels. Ideally, such an adaptive sequencing of tasks would explore hypotheses about the student's abilities based on early responses, and guide the student toward those tasks which have high probative value, a high likelihood of information return. Ad seriatim tests are seldom able to accomplish this. The "active onlooker," however, can take the form of a test specialist working one-on-one, or a computer algorithm suitable for real-time shaping of the testing task, working either rule-based or probabilistically. To do this requires both suitable heuristics on which to build adequate diagnostic inferences and a systematic way of moving rationally between test tasks.

A rule-based system, DX, functioning as a dedicated test administration/feedback device was programmed for use in conjunction with a test of reading specifically designed around diagnostic principles. The role of "active onlooker" was taken by a real-time interactive program called DX, described below. First, however, we present a review of the important concepts which drive the construction of a diagnostic test, and the results of two pilot studies conducted to evaluate and refine suitable test instruments.

Diagnostic test construction

The underlying goal in the development of the diagnostic tests was to create a profile of scores for each individual as well as for groups of students, which teachers can use to diagnose specific areas of difficulty. This is accomplished by 1) designing a test which rigorously measures those factors in a domain which can affect performance; and 2) determining the level and consistency of students' performance across items and item clusters which measure those factors.

A domain referenced approach was used to design test items for two separate diagnostic tests, the first in pronouns, a very well-bounded area in language skills, the second in more general abilities of reading comprehension. The approach that assumes that the main goal of testing is to assess an individual's status with respect to a skill requires a thorough understanding and specification of the domain to be addressed. The resulting domain specification provides a blueprint for developing test items, by way of a conceptual map of the skill to be assessed. The blueprint was drafted by reference to extant curricular material, subject

area specialists, and research on the structure of the knowledge base and the nature of learning. In addition, we identified important factors within the domain that might cause an item to be more or less difficult or a student's performance to vary. Items representing these factors were used to produce a test with diagnostic utility, one which identifies the reasons for the student's performance level.

PILOT STUDY: PRONOUN TEST

Within the framework of domain referenced testing, a diagnostic test was developed to examine pronoun use by fourth-sixth graders. This test was used to establish evidence for manipulating a number of variables within the test structure. Variables included four factors representing content structure and one representing cognitive complexity. At least in theory, each pronoun type could be classified by form, number, and person. There are two types of form: relative form (who or whom) and non-relative form. Number pertains to singular (she) and plural (they). Person can be of three types: first (I, we), second (you), and third (he, she, they). Since items measuring the second person would have sounded contrived to the reader, the test included only the first and third persons.

Two levels of cognitive complexity were used in this test, corresponding to whether students had to use the context of a reading passage to determine the correct pronoun. In the first level, the pronoun referent was given and the student needed only to associate that referent with the correct pronoun. In the second, more complex level, students were presented with a short paragraph that included a blank in the place of one noun; students needed to use the context of the paragraph to identify the

referent that was appropriate to the blank and then select the correct pronoun for that referent. The correct pronoun could be determined only from elements of the paragraph in which the pronoun was embedded. Consequently, the test used two levels of embeddedness corresponding to two levels of cognitive complexity.

The ideal Pronoun test would have items for every combination of the five factors. Since the form, embeddedness, person, and the number factors each had two levels and the rule factor had five levels, a complete test would have 80 (2x2x2x2x5) combinations. However, for several combinations of factors, sensible items could not be written. First, non-embedded items could not be written to elicit singular first person pronouns (I, me, or my). Second, items testing the relative form of first-person pronouns would have been contrived. Third, English does not use any relative form of possessive pronouns.

The test used a multiple choice format with five alternatives per item, consisting of the correct response, three distractors which were correct in all ways but one, and a fourth distractor which was correct only in one way or not at all. An example is the item, "Mom praised Mary and Stevie", with the following alternatives: them, they, us, him and she. The correct response (them) is an objective, plural third-person pronoun. The next three responses (they, us, and him) were correct on two of the next three factors (rule, number, person). The final response (she) was correct only in the person. The last response was considered a "wild card" distractor (a highly unlikely selection), included to detect guessing or carelessness.

Test Administration

Sample. Sixth-grade students from three elementary schools within a local inner-city district were involved in this study. These schools are located in a low to middle SES area with a high rate transient and mixed population. Approximately 90% of the students were of Hispanic background, 6% were Black, 2% were Asian, and 2% were non-minority Whites. There were 79 students classified as FEP (Fluent English Proficient) and 49 classified as LEP (Limited English Proficient), based on district reclassification criteria of language proficiency tests, achievement tests, and teacher judgments.

Procedure. Two forms of the diagnostic Pronoun test were prepared. Both contained the same items but the order of the items was inverted. After pilot administrations and feedback from teachers and students, the 92-item diagnostic test for pronouns was administered by project staff to 128 pupils. Test instructions allowed the administrators to clarify the meaning in vocabulary item stems but not in item distractors. Students were allowed up to 90 minutes to complete the test although most students finished the test in under 60 minutes. Classroom teachers were present during testing.

RESULTS

Performance on the Pronoun test was analysed using generalizability theory, a measurement theory designed to assess multiple sources of variation in a measurement . Generalizability analysis is particularly suitable to answering whether all students have difficulty with the same material (for example, all students misunderstand how to use relative

pronouns, or all students have difficulty with the sequence questions for expository passages). If true, then a single profile for the whole class may suffice for diagnosing areas of difficulty. If some material is particularly troublesome to some students and not to others, then profiles for individual students may be necessary.

Generalizability analysis also indicates if students perform equally well on a cluster of dimensions (such as all nominative, objective, and possessive pronouns; or all of the narrative passages). If so, then it would not be necessary to provide separate scores for each of those dimensions. On the other hand, if mastery of one dimension (such as nominative pronouns or passages which are narrative and contain much explicit information) is much greater than mastery of other dimensions (such as relative pronouns or expository passages), then it would be necessary to profile separate scores for each dimension. Additionally, this analytic technique indicates the number of items that are needed to reliably measure each skill presented in the profile.

Preliminary analyses examined whether there were distinct population subgroups in the design. Analyses of variance indicated that the only population characteristic influencing performance was language background (FEP vs LEP; $F = 30.09$, $p < .001$). The statistical tests for classroom, school, ethnic background, and age were not significant. Only the distinction between FEP and LEP was maintained in subsequent analyses.

The greatest sources of variance were due to pronoun form (relative vs non-relative); context (embedded vs non-embedded) and the pronoun usage rule (whether the pronoun is a direct or indirect object, for example).

PILOT STUDY: READING COMPREHENSION TEST

An extensive review of the literature on models of reading comprehension was conducted. Teachers and current texts for reading comprehension instruction were consulted to determine the extent to which test skills and variables were considered in the practical context of instruction. We concluded that the diagnostic assessment of reading should consider not only the comprehension skills to be assessed, but also the attributes of the text on which the assessment is based. Thus we selected the five most comprehensive models of reading; table 1 shows those reading comprehension skills and text variables identified as critical by the models. We reviewed the literature to determine which of these skills and variables had been well researched and what methodologies had been employed. We then selected those skills and text variables which had been most widely and vigorously researched.

Our literature review indicated that students' performance varies significantly in relation to the particular attributes of the text they read. Both novice and expert readers find some kinds of text, such as science passages written in the expository genre, more difficult than other kinds of text, such as fables. However, most of the studies examined only one or two features, such as syntactic complexity, degree of implicit information, or genre. Few examined a complex of features. We suspected that the degree of text difficulty may depend on the combination of features rather than on any single feature. Thus the passages used in the Reading Comprehension test represent all of the combinations of three major text features: genre (expository vs. narrative); syntactic complexity

(many vs. few subordinate clauses); and the degree to which information is explicitly or implicitly expressed. The curricular review yielded content and passages which were adapted for the test.

Student performance is also affected by the cognitive complexity of the processing of text. This is determined, in part, by the kinds of instructions, tasks, or questions readers are asked to attend to when reading text. These fulfill at least three functions. They set the reading goal for the reader; the amount and nature of information the reader must process; and how the reader needs to process the text (recall, synthesize, integrate, analyze, apply, etc.). The curricular and literature review indicated the following kinds of reading tasks/questions have been either researched or appear frequently in the curricula: literal comprehension questions, inference questions, main idea questions, and sequence questions.

To investigate the impact of each combination of comprehension skill/text attributes on test performance, items were generated for as many of the combinations as possible. For each combination, two parallel items were written. To the extent possible, parallel items included texts which were of the same genre; contained similar content, syntactical structure, and degree of implicit and explict information.

All passages were accompanied by the same categories of questions: one surface main idea, one underlying main idea, two literal comprehension, and two inference questions. Sequence questions were written only for the expository passages in order to avoid a test which was too lengthy and because the literature indicated that expository passages seem to be more difficult for expert as well as novice readers.

The narrative passages were based on Aesop's fables both in content and in their general story grammar. Such narratives were frequently found in the basal readers and presented a distinct, easily replicated structure. The expository passages were adaptations of science passages found in basal readers. These passages described either a process, such as metamorphosis, or a procedure, such as conducting a simple experiment.

The test used a multiple choice format with five alternatives per item, consisting of the correct response, three partially correct distractors and a distractor which was not at all correct. The final draft of the Reading Comprehension test, containing 20 passages and 136 items, was piloted in paper-and pencil format.

Test Administration

Sample. The sample consisted of fourth, fifth, and sixth graders of varied ethnicity, including Blacks, Hispanics, and Asians as well as non-minority students. Two-thirds were located in middle to high income areas within Los Angeles; one-third resided in a low to middle income area. All students were classified as either native speakers of English or as Fluent English Proficient.

Procedure. Two forms of the test, each containing the same items but in different sequence, were administered by staff. This was not a timed test. Most students, however, completed the test within 60 to 90 minutes. Students were permitted to ask questions regarding the pronunciation of words as well as the meaning of vocabulary which were not part of the item distractors or which constituted part of an answer.

RESULTS

Analyses examined the means and t tests for each item and for each combination of factors internal consistency across items; part-whole correlations for items within a passage and analysis of variance. Results from these analyses were used to determine which items were to be deleted or rewritten, and to identify levels of difficulty, either by passage or question type.

First, passages were ranked according to their level of difficulty, as represented by the means across the cluster of items which accompanied them. Second, each cluster of items accompanying a passage was examined to determine whether the items fell into a reasonable range of means (i.e., within a ten point spread) and whether the part-whole correlation of items within a passage was significant (.60 or above). If an item proved to be an outlier because of a low correlation or because the item mean was more than ten points above or below the means of other items in the cluster, the item was examined for possible flaws such as poor distractors and rewritten.

These results also indicated whether the items and passages fell into a hierarchical structure. Analysis of variance was performed to detect interactions or main effects by the content or cognitive complexity factors. These analyses yielded a consistent pattern for passage type difficulty (the content factor) across all samples. There were no interactions or main effects for the syntactic complexity variable.

The analysis yielded inconsistent and sporadic results regarding the effects of question type. The only consistency found was that sequence questions proved to be significantly more difficult than other question

types for science passages. Surprisingly, literal comprehension questions often proved to be more difficult than inference or main idea questions.

FIELD TRIALS

Test materials

Once the Pronoun and Reading Comprehension Tests were constructed and evaluated, their items were made the basis of shorter tests suitable for delivery by computer. The computer version of the Pronoun and Reading Comprehension Diagnostic Tests was designed to examine three levels of difficulty. The levels of difficulty were determined by item means and hierarchies indicated by the pilot test results. For both tests the levels as determined by item difficulty were as follows:

Level 1 (least difficult): mean range of .75 to .90

Level 2 (medium difficulty): mean range of .60 to .74

Level 3 (most difficult): mean range of .45 to .59

Constraints of testing time and programming complexity made it necessary to select only a subset of items from the paper and pencil test versions of the pronoun and comprehension tests. Thus only eighteen Pronoun Test items and six Reading Comprehension Test passages, each with three items, were used for the computer adaptation. However, the original visual layout of the paper and pencil test was maintained.

The selection of the Reading Comprehension Test passages was based on the overall mean for passage difficulty; the means per item; and the part-whole correlations for items within a passage. The overall mean also had to fall within the levels described above, as well as the items within a passage. The part-whole correlations were used to make sure that no

outliers were included. We also considered passage characteristics in this selection.

Software

A micro-computer-based test delivery system was constructed using Pascal, a language which allows for very efficient code, access to nonsequential input files, presentation of screen windows and other operational advantages over competing microcomputer languages. The operating environment was USCD p-system; machine response time under this system is very fast. The flow of testing was constructed according to the following rules:

1) Initial testing begins at the first item representing a middle level of difficulty. No less than four items at any level are administered in sequence.
2) Four correct responses within a level cause the student to be "moved up" to items at a higher level of difficulty; four wrong responses within a level cause the student to be "moved down" to items at a lower level of difficulty.
3) Testing terminates when the item pool is exhausted, or when the next available item difficulty level has already been used, or when four correct responses occur within the highest available level of items, or when four wrong responses occur within the lowest available level of items.

The algorithms which encode these rules actually comprise only 10% of the total Pascal code. The remainder is dedicated to screen management, file management, and item sequencing. To the extent possible, the software

was designed to be simultaneously -- and without contradiction -- user-friendly, robust, and generic. Within certain constraints any properly-designed diagnostic test could be delivered by this system.

Since the design of both tests included in this study allowed several items to accompany a single item stem, the program must interpret instructions which allow it to match the various pieces of text to appear on screen together. For each item the computer must recognize a legitimate keyboard response (either "a" or "A" for the selection of choice number one) and prompt for a retry if an invalid key is struck. It must also keep count and exit from the retry command if the student insists on hitting an invalid key repeatedly. The computer must be equipped to properly file multiple attempts at the same exam by the same student, and have fail-safe devices available for preserving response data if disk space is full or power is lost. In technical terms, the software is fire-walled.

Additionally, Pascal is a language with a high degree of protocol uniformity, which insures that it can be transported between dissimilar equipment with a minimum of software maintenance. The present software has shown itself to be fully compatible with the Apple II+ and IIe, using two floppy disk drives and an 80-column card. One drive contains the master programs and Pascal operating system components, while the other holds the test input files and preserves student response data. About two dozen sets of test results can be filed per disk.

Test Administration

Sample. One hundred and sixteen students in grades 4 to 6 in seven urban schools were included in this study. Teachers and administrators

were asked to select only those chidren whose reading levels were not above grade level. Approximately half were of Hispanic background, 25% Black, 20% non-minority White, and the remainder Asian.

Procedure. Two 18-item computer-managed tests were administered to each student individually in quiet school library settings, using an Apple II+ or IIe. Instructions for use of the computer were made available by staff members as necessary; instructions for the test itself were delivered onscreen. The majority of these students had prior exposure to the Apple computer as part of their math-science curricula. Total testing time varied between 12 and 45 minutes per student, with a modal time of under 20 minutes. Classroom teachers were not present during the testing. Administration/scoring protocols were run as an intact set, in which the student was asked to provide his/her name and the answer to one simple trial query, then asked to respond with the appropriate key upon reading each question. Items on the Pronoun Test (Appendix A) occupy little of the screen, and can be read rapidly; stems and items of the Reading Comprehension Test (Appendix B) require most of a complete screen (80 columns x 24 lines) and must be read in detail. Screen window management allowed the student to realize that each new test item need not necessitate a rereading of the item stem material, but it was not erased until it was no longer useful. For half of the sample, responses were immediately followed for four seconds, by feedback as to the correct answer, all or part of the screen was erased as necessary. The other half of the samples received no feedback at any time as to their progress.

RESULTS

The analysis of the data gathered during the field trials of the DX system is divided into three parts. First we present summaries of student performance. We then present data regarding the performance of the testing software itself, and conclude with information about the diagnostic interpretations generated by the software.

Summary: Students

Each student faced no less than eight items and no more than twelve from either the Pronoun or Reading Comprehension tests, due to the processing rules employed by the test administrator portion of the software. At least four questions would be given from the middle level of difficulty, and the remainder would be drawn from the higher level if performance on the middle level warranted, otherwise from the lower level of item difficulty. On the pronoun test, 59% of all students moved to the higher level of item difficulty: on the comprehension test, only 36% of students moved higher. While modal performance is described by moving in the same direction on both tests, 31% of the students who moved up on the Pronoun test moved down on the Reading Comprehension test, while 9% of the students who moved down on the Pronoun test moved up on Reading Comprehension test.

Scores ranged from 0% to 100% correct in both tests, but only 2% achieved perfect scores on both. Table 2 presents average performance results by test level. The figures strongly suggest that there is assymmetry between the two tests in terms on difficulty levels: while the lower level items on the Pronoun Test more often are answered correctly

than the lower level items on the Reading Comprehension Test, the opposite holds for the higher level items. However, we note here that such disparities do not in themselves speak to the success or failure of this approach to testing.

The current reading level for each student was requested from the classroom teacher. By conventional measures, 17% of the students were said to be functioning at or below grade 4, 44% at grade 5 and 39% at grade 6. Reading grade level is significantly related to test performance, as shown in Figure 1. If the student was in the lower reading grade level, his/her performance on average was up to 20% worse than his/her higher-level counterparts. This statement holds both for the upper and lower levels of test difficulty. Only a small percentage of those lower reading grade level students moved to the high levels of test difficulty. At the same time, a larger proportion of that group scored no more than one correct response in every third question. The number of students moving to the higher difficulty level of both tests is significantly related to the reading grade level ($F=3.49$, $p<.05$). Only 10% of those below the 5th grade level achieved this dual movement upwards; for those at the 5th grade level the figure is 27%, and for those at the 6th grade level, 52%.

In both the Pronoun and Reading Comprehension tests, the distractor patterns were revealing. The pronoun test produced a uniformity of errors -- approximately one third of all errors were committed in each of three distractor categories. Those students moving down committed more Nominative pronoun errors, a simple type of error, while those who moved up committed more Object-As-Preposition pronoun errors, a more complex error.

The Comprehension Test was not as balanced overall: errors of Main-Idea were not as frequent as Literal errors, the simpler error type, with the latter being made quite often by students who moved up. Table 3 shows the distribution of errors by error type within test.

As part of the field trial design, about half the students were given immediate feedback as to the correctness of their response or the correct answer if wrong. Those receiving feedback did no better or worse than those not receiving feedback, but in many cases, there was a palpable difference in attitude about the experience afterwards. Students in the feedback group were more likely to talk openly, and in some cases excitedly, about the experience they just completed. A high-scoring sixth-grader offered to consult with the programmer: "Whoever designed this did pretty good but he could use some of my help...!"

The first forty-two students who participated in this field trial were asked to respond to a simple 15-item questionnaire to detail their views of the testing experience. The overall impression from these responses is that the computer-managed testing seemed both easy and interesting. None indicated that the computer's instructions were confusing; few indicated they would be more comfortable taking the tests using paper and pencil. These students did not feel uncomfortable with computers in general, but the statement "It is just as easy to take a test on a computer as it is to take a paper-and-pencil test" elicited responses across the entire range, from strongly agree to strongly disagree. On the whole, the average respondent felt that the Pronoun and Reading Comprehension Tests were not difficult, and that the experience had been fun.

The single most frequent complaint from students was about the clarity of the computer screen display of the Apple computer. Standard school-owned equipment of various vintages and uneven degrees of focus was used in the field trials. Not under software control, this lack of screen clarity appeared to be a strain on several students and may have led to occasional inadvertant errors.

Summary Testing: Software

System performance was generally very good to excellent during this field trial. Perhaps the most dramatic failure occurred when a teacher came through the door looking for a spare computer power cord, and without thinking managed to "pull the plug" on a testing session in progress. Typing one's name and answering a mock test question proved to have a half dozen nontrivial variants. In other respects, most faults that could be predicted did indeed occur at least once, but the system included sufficient firewalling to allow testing to continue. Firewalling comprises those portions of software programming which enabled fault tolerance and system recovery under adverse conditions. For example, duplicate names and repeat sessions by the same person are assigned unique directory names. Fully one-sixth of the programming code is occupied by firewalling: happily, it generally lived up to expectations.

The system managed events in the testing session in several respects: it presented materials, captured and checked responses, and moved the student up or down from the middle range of item difficulty. After some starting problems were resolved, presentation of materials and checking of responses proceeded flawlessly; moving up or down must be evaluated

separately. In a finely-tuned testing system, such movement would be the result of complex probabilistic, heuristic, and/or logic-based real-time interpretations of student performance. The present system used a small packet of rules regarding response tallies, separately by test and by item difficulty level within test. Table 4 shows the results of the system's decisions with regard to sending a student to a more difficult or less difficult set of items. Test paths regarded as "correct" are those in which the number of errors made at the higher level is not less than those made below.

An example of an incorrect path is one in which a student scores moderately well on the middle level but fails every item on the lower level of difficulty. Unfortunately this illogical response pattern is entirely plausible, and indeed was present in 5% of the Pronoun Test responses, and 18% of the Reading Comprehension Test responses. The instances of apparent misfits were evenly divided between those students moving up and moving down. Fully 30% of students at the 4th grade reading level or less were apparently misfitted. Part of this may be due to random or nearly random responding by these students. However, during these field trials few students at any level were seen to hit the identical response key repeatedly and all appeared to be actively engaged in the testing task.

Summary: Diagnostic Interpretations

A relatively small number of error types were included in the test administered during the field trials. The plurality of errors on the Pronoun Test derived from confusion about use of the pronoun /whom/. One

Pronoun Test item showing this confusion had two distractors which each received more responses than the correct answer.

Several popular errors in the Reading Comprehension Test involved a distractor based on a key word found in the first line of the test passage. One distractor in each of four Comprehension Test items received at least as many responses as the correct answer.

In the context of the Pronoun Test, the distributions of answers alloted to each answer choice is highly uneven. One item (#18, the last of the lower difficulty items) was answered correctly by 88% of those who got to it. In contrast, one item (#2, the second item in the upper difficulty set) was answered correctly by only 26% of the students. In the Reading Comprehension Test, items ranged from 94% correct (#6, the last of the upper difficulty items) to 35% correct (#18, the last of the lower difficulty set).

These figures are at odds with the values found during the pilot studies, suggesting one of two conclusions. First, the fact of computerizing the tests may cause item level performance to be at variance to paper-and-pencil performance by virtue of some characteristic of the videoscreen or keyboard. Certain longer text items tended to fill the entire screen, and in some cases that resulted in pincushion or marginal distortions and therefore a reduction in clarity. Alternatively, the test items themselves may not be sufficiently robust in application to fifth and sixth grade pupils, computerization aside.

DISCUSSION

Diagnostic Interpretation

In practice, students were wholly unaware of the underpinnings of the testing, only that each test item was followed in order by another. Students did not seem to be concerned that testing times were often very different, or that what might appear or their neighbor's video screen would not appear on their own. In the present study, such optimizing is shown by the ratio of successful to unsuccessful test paths (using the definition provided earlier). For the Pronoun Test, the simpler of the two diagnostic tests, the ratio is good. For the Comprehension Test, which by definition involves substantially more ambiguity in both the nature of the test stimuli and the diagnostic interpretability of competing responses, the ratio is somewhat reduced. Obviously, a study using a less constrained set of test materials and a larger number of diagnostic indicators, while conceptually more accurate, is not likely to experience the same rates of sucessful test path management. The distinctiveness of "successful" vs. "unsuccessful" test paths in that study would be reduced unless all other variables are held equal. If the test domain itself is made less ambiguous, or if the strength of the diagnostic indicators is improved, improvements in test path susccess are possbile though by no means certain.

In contrast to ad seriatim testing, the controlled environment of a computer-managed testing session theoretically allows optimizing of the test to the emerging "picture" of the proficiency of the student. Several students asked to see their total scores, and, while disappointed to discover that total correct score was probably misleading, took an interest in the diagnostic particulars. Some then volunteered that they indeed recognized one or another error as a tendency in their classroom work.

The constraints to diagnostic interpretation are numerous and varied. In most extant work with computerized diagnosis, the system integrates a sizeable database of known diagnostic indicators, and constantly refers to that database as part of its deductive strategy. The present formulation relies instead on an abductive approach, which in turn relies on the strength of each test item and its relation to the targets of concern, and the probative validity of each response in terms of a given diagnostic hypothesis. Any item, no matter how well-formed and well-validated in pilot studies, may nonetheless elicit a unique response from an examinee. Subsequent diagnostic inferencing which fails to see that uniqueness will suffer accordingly. In the present study, unfortunately, we have no scientific means of making such assessment.

Student Behaviors.

From the behaviors of students involved in this study, it is clear that most upper primary pupils not only have some familiarity with microcomputers, but that their expectations of software are very high. This is no doubt due to the visual sophistication of most video arcade machines. Students waiting at a computer before a test session frequently made some effort to run BASIC or LOGO, only to be rebuffed by the unfamiliar p-system operating environment. Some of the brighter students appeared mildly frustrated that the actual requirements of the task were simplistic even if the test items were tough: only single keystrokes were needed to make one's response. Often, early in testing, the brighter

examinees would begin to type their answers directly, despite instructions on screen about hitting one key only. In sum, the visual excitement of a text-based testing system is low.

Limitations

In the present study three distinct lessons were learned regarding the limitations of diagnostic testing using computers. First, it is essential that all facets of the testing software be firewalled, for the weakest point in the software defines the system's weakness in actual operation. Unexpected keyboard entries and carriage returns, disk-swapping, directory errors, and other events even with low intrinsic probability are plausible and must be protected against.

Second, the number of test items needed to sufficiently explore competing diagnostic hypotheses is an exponential rather than linear function. A stream of correct responses interrupted solely by errors of a single kind is unlikely. Far more prevalent are response patterns which are due to one or more of the following: random guessing, inadvertant keystrokes, partial elimination of distractors, competing but nonexclusive diagnostic behaviors, and behaviors which are not strongly defined by any of the working pool of diagnoses under consideration.

Third, the nature of computerized management of testing in itself does not add a guarantee about the adequacy and appropriateness of the concluding diagnostic interpretation. Rather, it allows a more efficient route to that conclusion, which then must be appraised in the appropriate context of the theory or theories of reading behavior underlying the test itself.

The rule structure which governs event management of a testing session is critical to the utility of computer-aided diagnostic testing. In the present system, the decision to keep the rule structure to three levels of item difficulty (a functional minimum) was made both because the structure of the tests to be administered was strictly limited, and because adding additional complexity would not have added additional information for research purposes. Though the rules utilized in this field study make relatively few demands on the test designer and the programmer, they do assume that the examinee will answer items in a stable manner -- that is, answers will be consistent, items need not be repeated, one wrong response contains the same information value as another, and performance will not change materially during the testing session even though the student may come to feel familiar, bored or frustrated with the task.

In theory, each planned increase in rule structuring simultaneously should show an increasing well-tempered fit to diagnostic realities, and should enable a more flexible and adaptable system overall. One needed improvement is to build a system which selects the sequence of items with direct reference to the nature of the wrong answers given; answer A leads to items which specifically probe that response in detail. The granularity of diagnostic information derived from such detailed structuring is finer than before, though it risks cutting the analytic knife more closely than can be supported by theory. It must be noted, too, that rules which engender any ambiguity whatsoever in their real-time execution will cause significant problems for the software. What route should be formulated for the examinee whose performance improves or deteriorates over the course of

the testing session? Should a student who initially does badly always be relegated to a lower level of item difficulty? Alternatively, how much complexity in the specification of branching is likely to be applicable to the modal case? As is frequently true in real-time software design, major amounts of programming logic are required for minor amounts of operational enhancement, and/or for plausible but rare events.

Another issue which cannot be lightly dismissed is found in the heated, highly technical controversy about approximate and plausible reasoning, and/or endorsement or belief-based logic as opposed to conventional syllogistic or Bayesian probabilistic reasoning. In many respects, the multiple uncertainties surrounding examinee performance may be more appropriately modeled by fuzzy truth values and fuzzy proximities ("A is very much like B"; "C is fairly close to D") than by exact logics ("A is equivalent to B"; "C does not equal D"). These complex approaches to the heuristics of uncertainty, however well-suited in theory to diagnostic analysis of performance, are neither tractable for small samples nor easy to corroborate within the scope of educational testing. It is a case in which the available computational power potentially exceeds our present ability to use it well. One is forced to take a relatively conservative stance in the present circumstances, and use only such complexity as is warranted by the immediate task of providing rough diagnostic indicators.

Table 1

Reading comprehension skills and text variables by model

Table 1

Reading Comprehension Skills and Text Variables by Model

	Studies	General Trends/Issues
Automaticity Decoding	Calfee & Piontkowski, 1981; Lesgold & Resnick, 1982; Fleisher, Jenkins, & Pany, 1979	Analyzed 1st grader's acquisition of decoding causal links between automation and comprehension. Training in single word decoding improves comprehension.
Sightword Recognition	*Drum, Calfee, & Cook, 1981 Perfetti & Hogaboam, 1975	Examine effects of surface structure on comprehension. Relationship between single word decoding and comprehension: skilled readers more rapid. However, comprehension differences not clearly controlled.
Word Meaning (Vocabulary)	Beck, Perfetti, & McKeown	Relationship between rapid word access and comp: better comprehenders more accurate though not necessarily more rapid.
Processing Sentence Level Syntax	Marshall & Glock, 1978-79 Trebasso, 1980 *Barnitz, 1980	Logical relationships are signaled by connectors. Paragraphs with short sentences may be harder to comprehend than those with longer but more connected sentences. Lexical ambiguity, pronominal references, sentence context and comprehension. Effect of referents in 4 syntactic functions on comprehension, grades 2,4,6.

* Indicates studies with information which is useful for test development e.g., sample items, or domain specifications.

Theoretical and Historical	Studies	General Trends/Issues
Background Knowledge/ Schema	*Beck, Omanson, McKeown, Graves, & Cooke, 1981	Knowledge-expanding lesson on general content of a text (before subjects read passages) increases their comprehension.
	Ausubel, 1963	Advance organizers improve comprehension.
Content Knowledge	*Mossental, 1979 Pearson, Hansen, & Gordon, 1979	Hi vs. Low content knowlege and comprehension: high content knowledge increases comprehension.
&	see also: Langer, Nicholich, 1970;	Prereading activities aimed at increasing concept knowledge increases comprehension of low ability upper elementary and junior high students.
World View	*Graves & Cooke, 1981	
	Anderson, 1977 Lee & Allen, 1963 Stauffer, 1970	Using experience improves comprehension.

Note - Semantic and conceptual mapping are techniques used to discover or enrich subjects background knowledge throughout these studies.

Theoretical and Historical	Studies	General Trends/Issues
Text Structure		
Macro Story Grammar	*Collins, Brown, Larkin, 1979	People create & revise hypotheses as they read. There are micro as well as macro cues which signal hypotheses changes.
Passage Length	Newsome & Gaite, 1971	Short passages are recalled for a longer period of time, regardless of background information on the passage - see limitations.
Other, e.g., Essay Grammar		
Micro Propositions	Kintsch & Keenan, 1973;	Texts with greater number of propositions more difficult to comprehend. Texts with logically inconsistent texts propositions harder to comprehend.
	*Kintsch Kozminsky, et al, 1975;	A richly detailed text which develops only a few concepts is more comprehensible than text which develops several concepts - regardless of passage length. Propositions found in superordinate propositions will be recalled more easily than subordinate propositions. Limitation - no relationships drawn to referents.
	Baker, 1979	

Theoretical and Historical	Studies	General Trends/Issues
Social Interaction		
Reader's Perceived Purpose for reading	Gallimore et. al, 1982	Classroom context/culture and reading: Relationship of reader's home learning and communication style to classroom communication and learning with reference to reading. Includes references to building schema, and story grammars.
Prior Instructions/ Questions given to readers	Mosenthal, 1983 Langer & Nicholich, 1980	Metamemory, social context and comprehension. Effects of testing and prereading activity to help student access relevant information. Prereading activities improved comprehension for average readers only. No effect for low and high comprehenders.
Motivation & affect		

Theoretical and Historical	Studies	General Trends/Issues
Metacognitive Skills Awareness of non-comprehension	Baker, 1979 Markman, 1977	Readers use several strategies to cope (good & bad readers have developmental differences in their coping strategies). Bad readers stick to bottom up relevant strategies (i.e., decoding, sentence level syntax). Good readers use a variety of strategies including top down. (i.e., conceptual mapping, drawing analogies).
Strategies for for coping with Comprehension Difficulties	Clay, 1972 Olshavski, 1976-77	I.D. 1st graders' awareness of coping strategies. I.D. 10 strategies used by high school subjects for tackling difficult passages: examples: 1) concentrating on decoding and sentence level comprehension; using sentence level, paragraph level context for meaning; rereading beginning and reading end of passage to understand something in the middle.
General Strategies for approaching text	*Fredericksen, 1983 Anderson et al (T) Collins, Brown & Larkin	Reading as problem solving (see also Baird, 1982). Passages can pose ill-or-well structured processes. Identified 12 variables that may affect performance in finding the main point. Manipulatd 4 variables: fit of passage to strategy; how main point stated; amount of irrelevant information; frequency of supportive statements. Identified 8 problem solving strategies for approaching text. These strategies combine micro and macro structure elements. Overlap in strategies identified above.

Theoretical and Historical	Studies	General Trends/Issues
Testing	Drum, Calfee & Cook, 1981	Effects of surface structure variables on performance.
	Haertel & Calfee, 1983	Features for describing differences among reading achievement tests.
	Bauman, J.	Linguistic structure and validity of reading comprehension tests.
	Anderson et al	Domain-referenced approach to testing comprehension.
	Royer & Cunningham	Most R.C. test world knowledge rather than comprehension.
Hierarchies	Davis, 1968,72 Spearritt, 1972 Thornoike, 1973 John Carrol, 1976	Identified 5 skills; 4 factors or reading comp. Reanalyzed Davis' Data. Spearrit found 4 skills, 3 factors. Thorndike - 3 factors, also a reanalysis of Davis. Identifies 8 components of reading comp - theoretical only. Example of factors found throghout studies. - recall of word meaning - using context for meaning - inferences from content - recognizing author's purpose, attitude, tone - following structure (sequence) of passage

Table 2

Average Performance by Test Level

Test level	Pronoun		Comprehension	
	%correct	n	%correct	n
higher difficulty	49.5	69	71.8	42
middle difficulty	63.5	116	59.0	116
lower difficulty	77.8	47	60.2	74

Table 3

Errors by Error Type

PRONOUN:	Nominative	Direct Object	Object as Preposition
% of total errors:			
overall	31%	35%	34%
% of total errors:			
if S moved up	30%	34%	36%
if S moved down	38%	34%	29%

COMPREHENSION:	Main Idea	Inference	Literal
% of total errors			
overall	28%	36%	39%
% of total errors			
if S moved up	15%	30%	55%
if S moved down	30%	37%	34%

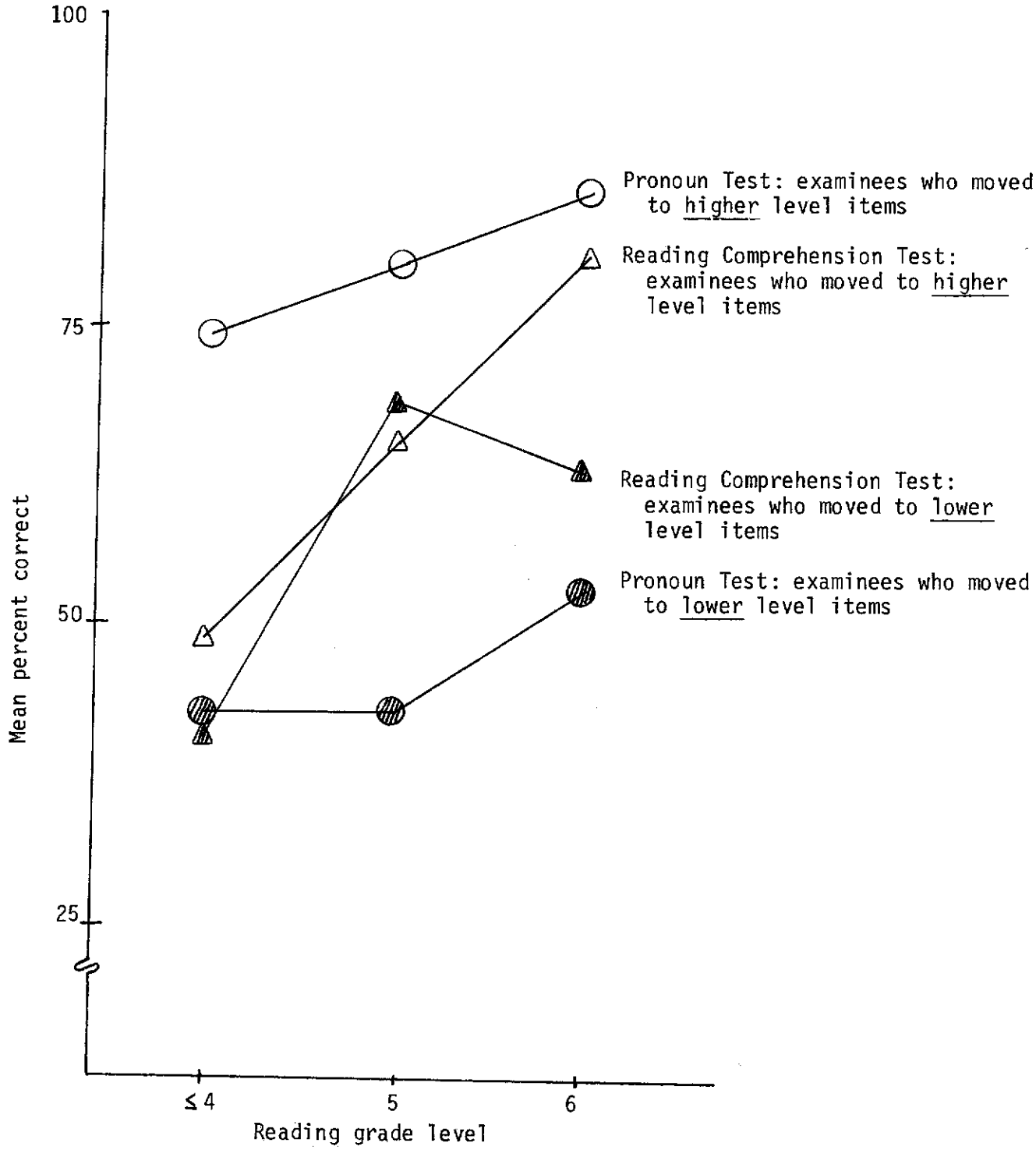
Table 4

Correct vs. Incorrect Test Paths

	PRONOUN	COMPREHENSION
"Correctly" moved examinee up or down	89%	70%
Apparent misfitted path for examinee	5%	18%
Unclear for results gathered	6%	12%
Ratio -- Moved up : Moved down within test	9:5	2:5
Moved examinee down on pronoun but up on comprehension	9%	
Moved examinee up on pronoun but down on comprehension	31%	

Figure 1

Reading grade level and test performance



APPENDIX A

The sentences below have a missing word.
Please select the appropriate pronoun to
fill in the missing words. Be sure to
keep the meaning of the sentence the same.....

The politician cleared his throat and looked solemnly at his audience. He began his speech by saying, "I want to express my appreciation to the volunteers, to ----- I will always be grateful."
a. who
b. they
c. which
d. her
e. whom

The professor was listening carefully to the famous astronaut. The astronaut, ----- the professor admired, had some interesting theories about the effects of space on the body.
a. whom
b. he
c. she
d. who
e. which

The campers were awakening to a crisp morning. John, ---- was usually the first one to get up, was snoring loudly.
a. she
b. they
c. which
d. whom
e. who

The actress ran to the stage and grabbed her award. She turned to the audience and said, "I want to dedicate this award to my parents, to ----- I will be eternally grateful."
a. who
b. they
c. him
d. which
e. whom

The children, ----- were excited about the spelling bee, were also looking forward to their surprise. Mr. Hodges had provided cookies and punch as a reward for their hard work.
a. them
b. which
c. who
d. whom
e. him

The boys were waiting for Mr. Jones, the vice-principal. The boys, ----- Mr. Jones had warned several times, were nervous about seeing him again.
a. who
b. which
c. whom
d. they
e. she

The sentences below have a missing word.
Please select the appropriate pronoun to
fill in the missing words. Be sure to
keep the meaning of the sentence the same.....

Jenny and her sister saw Tom, their brother, outside while they were cleaning the house. "Tom!" exclaimed Jenny, "Please come and help ----- girls clean up."
a. them
b. she
c. we
d. us
e. their

Mike and Larry were vacationing in London when they spotted the lead singer of a rock group, Clash. As they ran toward him, the singer ran away from ----- and disappeared.
a. them
b. their
c. us
d. him
e. they

Our team won the game by 10 points. However, ----- wanted to be good sportsmen, so we treated the other team to a pizza lunch.
a. I
b. he
c. we
d. they
e. them

Jim and I won the first place prize for our art. ----- wanted to celebrate our victory, so we all went out for soft drinks.
a. I
b. he
c. we
d. they
e. them

Sparkles was panting nervously as her owners tried to calm ----- down. She was about to have puppies.
a. him
b. she
c. her
d. them
e. he

The boss, with ----- Jan had to work, had a terrible temper.
a. he
b. him
c. whom
d. who
e. which

The sentences below have an underlined word or words. Please select the appropriate pronoun to replace the underlined words. Be sure to keep the meaning of the sentence the same.....

Mom praised MARY AND STEVIE.
a. them
b. they
c. us
d. she
e. him

The baby threw the ball to SARA.
a. her
b. him
c. she
d. his
e. them

JOHN AND MARY went sailing.
a. we
b. their
c. them
d. he
e. they

The horse galloped toward JOHN AND ME as we jumped over the fence.
a. them
b. us
c. him
d. we
e. theirs

MARY AND I went shopping together.
a. he
b. we
c. they
d. I
e. us

The police made THE BOY go home.
a. he
b. her
c. them
d. his
e. him

APPENDIX B

There are many active volcanoes on earth such as the famous Mauna Loa in Hawaii. They occasionally shoot smoke and lava onto the earth's surface. The formation of a volcano can be thought of as an underground balloon, buried under thin layers of sand and plaster. Hot magma fills the balloon and pushes up the ground. Magma is the hot material that forms the earth's center core. The magma pushes upward inside the volcano and melts the surrounding rock and dirt. The melted rock and dirt is called lava. Eventually, the lava will push through the top of the mountain. This action causes a volcanic explosion and forms a crater. After the explosion, or eruption, the lava cools off and the mountain sides shrink somewhat. This is similar to a balloon shrinking after it pops. A volcano is said to be inactive when it no longer produces eruptions.

This passage is mostly about how....
a. volcanoes are filled with lava.
b. mountains are sometimes volcanoes.
c. volcanoes are created.
d. magma creates mountains.

An active volcano...
a. erupts every other month.
b. erupts occasionally.
c. never erupts.
d. contains cold lava.

According to the passage, when a balloon pops, its sides...
a. shrink.
b. become larger.
c. crack.
d. become rounder.

Butterflies are created through an interesting series of stages. Adult butterflies lay their eggs on leaves or tree branches. These eggs turn into caterpillars. The caterpillars crawl around the trees and bushes, feeding on leaves. Eventually, the caterpillars weave themselves a silken shell called a cocoon. The caterpillars sleep inside the cocoon for a few weeks. Inside the cocoon, the caterpillar grows its butterfly wings, antennae and body. When it awakens, the butterfly breaks out of its cocoon and flies away. When it lays eggs, the process will begin again.

The passage is mostly about how....
a. adult butterflies are made.
b. caterpillars find food.
c. cocoons are created.
d. butterfly eggs are hatched.

According to the passage, how does the caterpillar use the cocoon?
a. The caterpillar stores its food in the cocoon.
b. The caterpillar eats the cocoon while it changes into a butterfly.
c. The caterpillar sleeps in the cocoon while it changes into a butterfly.
d. The caterpillar lays eggs in the cocoon where they hatch.

What do caterpillars eat?

a. eggs.
b. silk.
c. leaves.
d. branches.

A crow who had stolen a piece of cheese was flying toward a tall tree. She was going to eat her prize, when a hungry fox saw her. "If I plan this right," thought the Fox, "I'll have cheese for supper."

So as he sat under the tree. Fox falsely spoke in the most polite tones, "Hello, Miss Crow, how pretty you look today!" he lied. "Your wings are so broad and your feathers are so black. You hunt so well. I haven't heard your voice but I'm sure that it's finer than that of any other bird."

Crow was so flattered that she waggled her tail and flapped her wings to show her pleasure. She really liked what Fox said about her voice because she had been told that her caw was a bit rusty. So, laughing inwardly, she decided to surprise the Fox. She opened her mouth to sing with her beautiful voice. Down dropped the cheese and the Fox snatched it. Licking his chops Fox said to the Crow, "Next time someone praises your beauty be sure to hold your tongue!"

What did Fox say was pretty about Crow?

a. Her body, voice, and hunting ability.
b. Her tail.
c. Her feathers.
d. Her cheese, feathers, and singing ability.

What was Crow's surprise for the Fox?

a. Her cheese.
b. Her voice.
c. Her beauty.
d. Her wings.

This story is mostly about how...

a. the crow and the fox made friends.
b. the fox outsmarted the crow.
c. the fox lied.
d. the crow dropped the cheese.

How would you like to know how to make a steam engine in your kitchen? For your safety, be sure to try this experiment with an adult around. First, make a pinwheel with some paper and a thin stick. Cut the paper into a circle, or into two propellers. Now push a pin through the middle of the circle or propellers and pin it to the top of the stick.

Now put the stick on a clothespin. Next, get a teapot. Then put the clothespin on the teapot's spout. Now fill the teapot with water and put it on the stove. Let the water come to a boil. Now watch the steam make the pinwheel spin. The stream of steam coming out of the spout pushes the pinwheel and makes it spin. Steam is hot moist air. All hot air travels upward. Now turn off the teapot and watch the pinwheel come to a stop as the steam is no longer coming out.

This passage is mostly about...

a. how to make a steam engine from a teapot.
b. how to use steam carefully.
c. how to put together a pinwheel
d. how to use the stove safely.

In the passage, a pinwheel is a:

a. wheel of sticks held together by paper and a clothespin.
b. firework fixed on a stick which turns around when lit by a match.
c. lollipop which is the shape of a circle and has a spiral design on it.
d. wheel of paper attached by a pin to a stick so that it will turn around.

Why will the pinwheel stop when the teapot is turned off?

a. The steam is coming out.
b. The pinwheel is too soggy.
c. The steam is no longer coming out.
d. The pinwheel no longer works.

One day all the trees and flowers were gray. The great leader looked around and said, "This is dreary!" Who wants to paint all the flowers and trees for me!" Peacock volunteered to do it. In those days, he was ugly and had short tailfeathers, so he felt very bad abut himself.

The great leader told Peacock how to do this job. Peacock sadly said, "I'm so ugly. I don't know if I can do this job. What do I know about beauty? But the great leader made Peacock do it.

Peacock flew all over the earth. He collected all kinds of colors and used his tailfeathers to paint. At the end of the day, Peacock's tailfeathers were falling out.

"You have done a wonderful job," said the great leader. "It shows the beauty inside of you. I will give you long colorful tailfeathers to remind you of this."

This story is mostly about...

a. why the Peacock was ugly.
b. how the Peacock got its tailfeathers.
c. why the flowers were gray.
d. how the Peacock lost its tailfeathers.

Why did the great leader want the flowers and trees painted?

a. Everything was purple and he thought this looked funny.
b. Peacock collected all kinds of colors.
c. Peacock volunteered to paint them.
d. Everything was gray and this made evrything look drab.

Why did the great leader give the Peacock pretty tailfeathers?

a. To remind Peacock of his achievement.
b. Because Peacock's feathers were short.
c. To punish Peacock for his poor job.
d. Because Peacock was gray and ugly.

Along the shores of a lake lived a man named Parsee. He loved to eat cake. One day he made himself a giant cake three feet thick. Just as Parsee was going to eat the cake, along came a rhinoceros. In those days, Rhino had very smooth, tight fitting skin.

Rhino saw the cake, spiked it with his horn, and ate most of it. Parsee was very angry. Rhino was rude, he should have asked for a piece of cake.

A few weeks later there was a heatwave. It was so hot that everyone took off their skin and went swimming. Parsee found Rhino's skin and filled it with dried cake crumbs. He was going to teach Rhino to be polite. The crumbs would remind Rhino of what he had done.

When Rhino put his skin back on, the crumbs made him itch. The more he scratched the worse it itched. His skin has been baggy every since.

Why was Parsee angry at Rhino?

a. Because Parsee loved to eat cake.
b. Because Rhino was rude to steal the cake.
c. Because Rhino left his skin by the cake.
d. Because Parsee felt irritable in the heat.

What did Parsee want to teach Rhino?

a. To be neat.
b. To be polite.
c. To feel itchy.
d. To eat less.

This story is mostly about...

a. how much Rhinos love cake.
b. how the Rhino got baggy skin.
c. how Parsee got hungry.
d. how animals used to cool off.