Three Response Types for Broadening the Conception of Mathematical Problem Solving in Computerized-Adaptive Tests

Randy Elliot Bennett

Mary Morley

and

Dennis Quardt

Educational Testing Service

Princeton, NJ 08541

rbennett@ets.org

Paper presented at the annual conference of the National Council on Measurement in Education, San Diego, April 1998. Appreciation is expressed to Dave Bostain, Kevin Singley, Probal Tahbildar, and Alex Vasilev for their contributions toward creating the response types described in this paper. Funding was provided by the Graduate Record Examinations Board and by ETS.

Three Response Types for Broadening the Conception of Mathematical Problem Solving in Computerized-Adaptive Tests

Abstract

We describe three open-ended response types that should broaden the conception of mathematical problem solving embodied in computerized-adaptive admissions tests. The Mathematical Expressions response type presents single-best answer problems that call for an algebraic formalism, the correct rendition of which may take an infinite number of surface forms. Generating Examples presents loosely structured problems which may have many good answers, taking the form of a value, letter pattern, expression, equation, or list. Graphical Modeling problems ask the examinee to represent a given situation by plotting points on a grid, where there may be a single-best answer or multiple correct responses. Finally, these three basic types can be combined to form extended constructed-response problems. For each of the three basic types, this paper gives example items, describes the examinee interface and approach to automatic scoring, and reports available findings.

Current postsecondary admissions tests tend to take a relatively restricted view of the nature of mathematical problem solving. This view is evident in the questions the tests present. Those questions are primarily multiple choice and, as a result, invariably well-determined; that is, they are tightly structured, containing all the information needed to reach a single best answer. Also, questions typically focus on formalisms, giving limited attention to more qualitative representations.

This restricted view of mathematical problem solving is replicated in the computerized versions of admissions tests like the GRE General Test and Graduate Management Admission Test. This replication is understandable as the priorities in first-generation computerized tests are on making the transition from paper to electronic delivery and from linear to adaptive testing. However, computerized testing offers an opportunity to improve assessment far beyond the conveniences associated with electronic delivery and adaptive testing (Bennett, 1997).

Our work is targeted toward broadening the view of mathematical problem solving underlying computerized adaptive tests. This broadening is based on a description of mathematical problem solving in terms of structure and representation. This description is closer to, but not as comprehensive as, conceptions currently favored in the mathematics education and research communities (e.g., NCTM, 1989).

Following Simon (1978), problems are asserted to lie on a continuum from well- to " ill" -structured. Ill-structured problems have more complex and less definite criteria for knowing when a solution is reached; do not include in the instructions all the information needed to answer and offer only a vague sense as to what information is relevant; and have no simple " legal move generator" for finding all of the alternative possibilities for each solution step. Many of the more important problems encountered in

academic disciplines, in job settings, and in everyday life are of this kind (Frederiksen, 1984).

Problem representation refers to the type of reasoning the examinee must use to model a given situation. While different representations may be used during intermediate problem-solving stages, admissions test questions tend to call for responses that take a quantitative form. This emphasis ignores the importance of qualitative representations. Research has documented qualitative reasoning as one of several dimensions along which experts and novices differ (Glaser, 1991). Proficient problem solvers are able to reason using both formal representations and less formal, sometimes intuitive, qualitative descriptions. In addition, competence in representing situations qualitatively is generally regarded as a sound basis for learning formalisms. Finally, the ability to translate among different representations of the same situation, or of the same mathematical concept, is regarded as a powerful problem-solving tool (NCTM, 1989, p. 146). Consequently, the qualitative aspect of mathematical reasoning may provide an important complement to the symbolic element we now emphasize.

Table 1 crosses problem structure and representation to offer an idealized, high-level view of mathematical problem solving. As noted, current admissions tests generally accent well-structured, multiple-choice questions that call, at least in terms of the final answer, for a quantitative representation. Such tests concentrate their coverage on the upper-left quadrant of Table 1. In this paper, we describe three automatically scorable response types for computerized-adaptive tests that, in combination, should move us toward a broader conception of mathematical problem solving. These response types extend coverage both within the upper-left quadrant and more fully into other sections of the table.

Insert Table 1 about here

What is a Response Type?

Bennett and Bejar (1997) propose a model for the components of a computer-based test that includes a construct definition, along with a test and task design that operationalizes the construct; an examinee interface; a tutorial; test development tools; an automated scoring routine; and some method for communicating assessment results. An extension of the paper-based notion of item type, the term " response type" is narrowly intended to subsume the examinee interface, the general task design or task class that this interface delivers, and the scoring routine. The response type exists as an off-the-shelf package available to assist in operationalizing some target construct. This operationalization occurs through a systemic interaction among the computer-based test components. A response type puts broad boundaries on the construct domain that can be measured, so that any given application may represent only a portion of the domain that could potentially be covered. More precise specification occurs through the test design and through finer levels of task description.

The response types we depict have several characteristics. First, as for all response types, ours are quite general. For each, items can be written that range widely in difficulty and in the skills emphasized. We try

to communicate this generality by example. However, we give the most emphasis to task classes covering skills that are arguably somewhat distinct from those that would be represented by conventional versions of these same types. Second, although we present our response types as a conceptual package, each is an off-the-shelf component. Testing programs will use (or not use) these components as they see fit. In all likelihood, our types will be integrated (individually or in combination) into tests comprised of conventional types. Third, each type is open ended in deference to the concerns expressed by educators and testing critics who feel that standardized assessment falls short because of the narrow substantive limits that dependence on the multiple-choice format enforces. Thus, our argument for these response types is based largely on the proposition that they move us toward a theoretical conception of mathematical problem solving more consonant with the values of the education community (e.g., NCTM, 1989), and that could not be realized solely through the dominant question format. Finally, these response types are automatically scorable in real time, making them feasible for use in any situation where immediate or inexpensive scoring is warranted.

Mathematical Expression

The Mathematical Expression (ME) response type allows presentation of any item for which the response is a rational symbolic expression (Bennett, Steffen, Singley, Morley, & Jacquemin, 1997). The response type was created primarily to present mathematical modeling problems as part of an experimental GRE test for applicants to quantitatively oriented graduate programs. For this test, the development committee felt that it was critical to include items requiring the construction (as opposed to recognition) of formal mathematical models.

Interface. Figure 1 illustrates the ME interface along with a sample problem. (Additional, lower-level, problems are given in Table 2.) The examinee enters expressions of any desired complexity by clicking on the buttons shown. Among other things, these buttons represent numbers and operators, bring up a variable and constants menu, and turn on superscripts or subscripts. There are several interface features designed to prevent entry of responses that, through input mistakes, might prove unscorable. These features include disabling some response sequences (e.g., entering a radical in superscript mode thereby creating an irrational expression), disallowing keyboard entry to discourage examinees from entering superfluous text, and checking syntax for such things as unbalanced parentheses or illegally juxtaposed operators.

Insert Table 2 about here

Insert Figure 1 about here

Automatic scoring. The challenge for scoring ME items is one of mathematical paraphrase. That is, there are infinite ways to express the same mathematical relationship. For example, in field trials, the following

were among the correct responses examinees produced for the problem shown in Figure 1:

-1/4 * x + (9/4)	1/4 * (9-x)
-x/4 + 9/4	(-x +9)/4
-.25x + 2.25	2 - 1/4 * (x-1)
(9-x)/4	(-1*x + 9)/4

To meet this challenge, two scoring programs were created, each using a different approach. The first program used well-established symbolic computation principles to test examinee responses for algebraic equivalence with the key (which is also a mathematical expression). For example, given the key, $x/2$, and an examinee entry of $2x/4$, the algorithm would construct the expression, $(x/2)-(2x/4)$. Next, the algorithm would reduce the $2x/4$ term by 2 and subtract the remainder from $x/2$. The result of 0 would imply that the key and response were algebraically equivalent.

The second approach to scoring employed an evaluation methodology. Again, the algorithm would use the key and examinee response to construct the expression, $(x/2)-(2x/4)$. It would next evaluate this expression at $x=0$ and finding a 0 result, conclude that both expressions had the same value at $x=0$. The algorithm would proceed to evaluate the constructed expression at $x=1$, again finding a 0 result, suggesting that both given expressions had the same value at $x=1$. Since if two points of one line (in this case, $y=x/2$) match two points of a second ($y=2x/4$), all points of the first will match all points of the second; therefor, no more points need be evaluated to conclude that the key and examinee response are equivalent. For nonlinear expressions, the number of points needed is one greater than the degree of the polynomial. Currently, 50 points are evaluated at random regardless of the degree of the response, providing more than enough accuracy for the kinds of questions that need to be scored.

The two scoring methods have notable differences. One difference is that the symbol manipulation approach is more theoretically sound and virtually guaranteed to produce correct results. A second difference is more practical. The symbolic engine performed well with the high-powered work stations it was developed on. However, since adaptive testing requires real-time scoring, the algorithm needed to work on the low-end machines that populated test centers. When we tested it on center machines, we found that it demanded more computing resources than those machines could provide, such that very complex expressions took too long to score. To prevent the slowness of scoring from interrupting the flow of testing, the symbolic engine was modified to return a correct or incorrect score within one second, otherwise marking the item response as " indeterminate."

To test scoring accuracy, Bennett, Steffen, Singley, Morley, and Jacquemin (1997) ran the symbolic engine against the responses of 1,864 volunteers, each of whom took four of 33 ME items along with questions of other types. Of 6,834 nonblank ME item responses, there were 26 instances where the algorithm marked as wrong what human review revealed to be right. This constituted an accuracy rate of 99.62% or 38 errors per 10,000 responses, approaching the rate for scanning multiple-choice answers which

runs about 99.95%, or 5 errors per 10,000 answers (J. McDonald, personal communication, July 18, 1996).

For ME, the 26 miss-scorings fell into two categories. Twenty-three involved examinees using subscripts incorrectly, causing the algorithm to misinterpret the response (e.g., entering *C1* instead of C_1 or entering r_o, with the subscript being a letter, instead of r_0 where the subscript is a numeral). The second category involved 3 otherwise correct entries where the examinee used the variable and constants menu to enter unit abbreviations (e.g., " mi" for miles). These abbreviations were treated as concatenated variables (e.g., *m* x *i*) and thus misinterpreted.

We have since rerun the evaluation engine against this same data set. It, too, scored all but the same 26 (of 6,834) responses accurately, thus matching this aspect of the symbolic engine's performance.

Whereas this data set was processed off-line, we made sure to do that processing on a low-end machine so that we could identify the frequency of unscored, indeterminate responses. As it turned out, the symbolic engine produced a score for all submissions, primarily because the items posed did not require complex responses. However, there is always the possibility that by mistake or malice, an examinee will enter such a response.

Because we wanted to push the limits of the symbolic engine, we had test developers concoct 20 hard expressions and 292 difficult paraphrases. The symbolic algorithm was able to process 205 (70%) of the paraphrases, all of which it scored accurately. The algorithm rejected the remaining 87 paraphrases (returning scores of " indeterminate"), because the expressions exceeded its processing limitations. When we reran the evaluation engine against these 20 difficult expressions (again on a low-end machine), it scored all 292 paraphrases correctly.

An outstanding issue is how to handle the response classes that both engines scored inaccurately. With multiple choice, scoring errors are caused primarily by improper examinee entries, such as failing to completely erase one darkened bubble after marking another. For that situation, the first-line remedy has been to train examinees through the test bulletin to avoid making such responses. For ME, the root cause appears also to be improper entries, some of which we can correct for computationally more easily than others. For instance, we can readily change the processing to treat all entries of subscript letter " o" as subscript zero, thereby removing one source of error. However, given the low frequency of the remaining errors and the difficulty associated with handling them computationally, we will try to reduce their occurrence through clearer test directions before attempting more demanding approaches.

Psychometric characteristics. Because ME was initially developed as part of an experimental test for applicants to quantitatively oriented graduate programs, the published psychometric work has focused on evaluating its functioning in that context. While these results represent only the early stages of a larger empirical effort, they are relevant to the validation argument.

For the experimental measure, several response types were field tested including standard multiple choice, entering numeric values, and shading

portions of a coordinate system. Results reported by Bennett et al. (1997) showed ME items to have roughly the same distribution of difficulty as questions from the other response types. In addition, ME questions had item-total correlations similar to those for other items. Third, ME items took no longer to answer than other constructed-response problems written to measure mathematical modeling skills (though both types took longer than multiple-choice modeling questions). Finally, ME showed gender differences comparable to those for other quantitative questions.

In comparison with the other response types employed in the experimental measure (and with those in adaptive tests generally), the ME interface is quite complex. That complexity is cause for concern as the individual differences we capture in ME performance could well reflect, in part, proficiency with the interface. To test that hypothesis, Gallagher and Cahalan (in progress) asked approximately 200 examinees to complete parallel computer-delivered and paper-and-pencil ME tests (where the latter version simply posed ME questions on paper and allowed the examinee to hand-write an expression). In addition to these two tasks, participants took a computerized ME editing and entry test that required them to compose or modify given expressions using the ME interface. Gallagher and Cahalan then attempted to predict computer-based ME scores from the editing and entry scores, after controlling for paper-based ME performance. Preliminary results failed to show any main effect for the editing and entry test.

While the evidence to date shows no negative effect of the ME interface on the scores of those intending to apply to quantitatively oriented graduate programs, we must continue to improve the design. For one, we might expect those with less computer familiarity to have greater difficulty expressing their mathematical knowledge than members of the current target population. Second, even those in our population have reported deficiencies. One such deficiency comes from the fact that examinees are accustomed to the speed and ease of expressing themselves mathematically on paper--and, in fact, work ME problems that way until entering their final answer. The temporal juxtaposition of working on paper and then computer makes the convenience differences between the modes all the more noticeable. Thus, it is no surprise that Gallagher and Cahalan report many examinees feeling that computer input was too slow. More worrisome, however, is that many participants also indicated the interface impeded solution entry. In fact, a review of the paper-and-pencil ME responses located answers that wouldn't have been accommodated easily by the computerized version (e.g., fractions whose numerators and denominators included other fractions, making a result that would need to be recast to fit the answer box). We are confident we can make design improvements to remedy this situation and, with advances in handwriting recognition, eventually provide an input mechanism that may approximate the ease of working with pencil and paper.

Generating Examples

As noted, the (modest) innovation of ME is that it allows us to present problems that have correct answers taking many different quantitative surface forms and score them automatically in real time. In terms of the framework presented in Table 1, ME adds to the range of problem types that can be presented within the " well-structured quantitative" classification.

As we have suggested, however, the problems within this classification represent only a segment of the universe of mathematical tasks people face in academic and work settings. To be sure, some portion of those tasks are more amorphous or " ill-structured" (Simon, 1978).

It is, unfortunately, very difficult to score ill-structured problems with standard computational methods, so such problems will not readily work in adaptive tests. All the same, we should be able to widen the conception of mathematical reasoning represented in adaptive tests by including tasks that, to differentiate them from Simon's (1978) conception, might be better called " under-determined." Our formulation, dubbed " Generating Examples," (GE) presents constraints and asks the test taker to pose one or more instances that meet those constraints, where for any given problem there may be many such instances that, in contrast to ME, are mathematically different. As for their ill-structured cousins, not enough information is given to fix the solution uniquely. But in contrast with ill-structured problems, there is enough information to know when the problem is solved and, as a result, Generating Examples questions can be automatically scored. Thus, GE provides a practical mechanism for pushing the conception of mathematical problem solving toward the " ill-structured quantitative" cell of Table 1.

Table 3 gives instances of GE items. These instances call for either the entry of values or, in the case of item #4, an expression. Note that other objects, including letter patterns and lists (e.g., of multiple values) could also be used.

How do GE items differ from conventional well-determined questions? Nhouyvanisvong, Katz, and Singley (1997) have proposed a framework for understanding the differences in terms of problem-solving strategy. Through cognitive task analysis, they identified three strategies: generate-and-test (useful for either problem type), formal algebraic (only useful for well-determined problems), and a hybrid (also useful for either problem type). Nhouyvanisvong et al. suggest that four cognitive processes are implicated in the " generate-and-test" approach. These processes are deriving explicit and implicit constraints, estimating potential solutions, propagating values through constraints, and updating the estimated solution. Deriving explicit and implicit constraints is also required in the formal algebraic strategy but once that is done, two processes not used in " generate-and-test" are required: combining constraints into equations and solving the final equation. The cognitive processes implicated in both the formal algebraic and generate-and-test approaches compose the hybrid strategy. Because the examinee generates an algebraic equation and then creates values to propagate through it, this strategy is essentially a " plug-in" method. In essence, the Nhouyvanisvong et al. analysis, suggests that by requiring either the generate-and-test or hybrid approaches, GE forces the examinee to bring to bear cognitive processes different from those needed to attack well-determined problems via the formal algebraic strategy commonly emphasized in school.

Insert Table 3 about here

Interface. Because GE items can be written to call for the entry of different objects, this item class uses more than one interface. For GE items calling for an expression, the ME interface can be used. Items requiring input of single or multiple values can employ established interfaces already used in operational computer-based tests (e.g., Praxis I) to deliver conventional numeric entry items. Figure 2 shows a GE item and interface allowing the entry of multiple values.

Insert Figure 2 about here

Automatic scoring. GE items are scored by creating in executable code a key that embodies the conditions an answer must meet to be correct. Responses are then tested against these conditions by a superset of the evaluation methodology used for ME.

Consider the following item:

If n and m are positive integers and 11n-7m=1, what are two different possible sets of values for n *and* m*?*

The conditions contained in the executable key would essentially specify that a creditable response:

- Contain two pairs of values,
- Have a second pair different from the first,
- Have each member of each pair be a positive integer,
- Return for the first pair a true result when its values are substituted for *n* and *m* in the equation, 11*n*-7*m*=1, and
- Return for the second pair a true result when its values are substituted for *n* and *m* in the equation, 11*n*-7*m*=1.

The clean separation of these conditions raises possibilities for partial credit scoring. For the moment, however, we have limited our analyses to dichotomous grading.

So far, we've done only preliminary work on the accuracy of GE scoring and only with item types other than ME (Bennett, Morley, Quardt, Singley, Katz, & Nhouyvanisvong, in press). We did this work using responses to two parallel 20-item tests. For these 40 items, we wrote automatic scoring keys based solely on the constraints contained in the item stems (i.e., without reference to the responses given by examinees). The items were taken by 257 volunteers who were either planning to apply to graduate school or were already first-year graduate students (a considerably more heterogeneous sample than the ones used to evaluate the ME response type). We randomly selected 21 examinees for each form, machine and hand-scored their responses, and compared the results. In all 840 cases (42 examinees x 20 responses), the machine and manually generated dichotomous scores agreed. However, we know that this result is slightly optimistic, at least with respect to GE items that use the ME interface. For these items, there is the possibility that examinees will occasionally formulate correct entries that can't be

meaningfully interpreted because of improper response entry (e.g., they include unit labels as text).

Psychometric characteristics. Bennett et al. (in press) evaluated the functioning of the GE response type in the sample mentioned above. The Generating Examples tests were reasonably reliable, with internal consistency estimates of .82 for each 20-item test. GE scores were moderately related to the GRE General Test quantitative section (r = .66) and had a pattern of relations with other GRE sections, UGPA, and self-reported accomplishments that was similar to the pattern shown by the quantitative section. Even so, GE's disattentuated correlation with GRE quantitative did not approach unity, suggesting the possibility that the two tests do not fully overlap. This possibility is in keeping with the Nhouyvanisvong et al. (1997) model which stipulates that, whereas there is some redudancy, GE does indeed evoke cognitive processes--such as estimating potential solutions and propagating values through constraints--that conventional mathematical questions often do not.

Bennett et al. (in press) used this cognitive model to pose and experimentally test several hypotheses about the determinants of GE item difficulty. One consequence of the fact that GE questions can have many right answers is that examinees can be required to give more than one response. Bennett et al. found difficulty to be affected markedly by requiring two pairs of examples instead of one, with those participants randomly assigned to the multiple-response condition scoring about .6 standard deviations lower than the group required to give one pair of examples to the same items. From the perspective of the Nhouyvanisvong et al. (1997) model, the effect might come from having to repeat solution processes, raising the chances that individuals with inadequate understanding will falter.

Bennett et al. also examined GE's impact on women, as a preliminary step in addressing fairness concerns. They found that men consistently scored higher than women by about the same amount as on the General Test's quantitative section. When they controlled for General Test scores, no significant effects remained, suggesting that GE would not likely increase gender differences beyond those already present on the General Test.

Examinee perceptions. What do examinees think of Generating Examples items? When asked for their perceptions of the Generating Examples response type, examinees were evenly divided (40% to 39%) over whether GE was a fairer indicator than multiple choice of their ability to succeed in graduate education (Bennett et al., in press). At the same time, the overwhelming majority (72%) said that they would prefer to take a multiple-choice test. Such apparently contradictory sentiments in examinees' perceptions of constructed-response vs. multiple-choice items have been reported elsewhere (e.g., Braswell & Kupin, 1993; Bridgeman, 1992). These sentiments may simply represent a view of testing as nothing more than a necessary evil; that is, taking a test that is easier to prepare for is better than taking a harder one more relevant to graduate study.

Graphical Modeling

Both ME and GE encourage examinees to deal with problem situations quantitatively (at least at the level of the final answer). But cognitive

research, as well as current notions of what it is important for students to know and be able to do in mathematics, suggest that greater attention be paid to the use of qualitative representations in education and assessment (Glaser, 1991; NCTM, 1989, p. 146). Qualitative representation might involve verbally describing the problem situation, rendering it diagrammatically, or modeling it graphically.

Table 4 presents some illustrative Graphical Modeling (GM) problems. Note that such problems may depict either under-determined or well-determined situations (putting GM within the " qualitative" row of Table 1.) To illustrate the generality of the problem type, examples #4 and #5 are cast in pure contexts that do not require qualitative modeling per se but do require the examinee to demonstrate mathematical concepts graphically. Problem #4, in particular, raises the possibility of testing geometry content more directly than could be achieved with conventional item formats.

Insert Table 4 about here

Interface. Figure 3 gives the GM interface. As indicated, each problem is accompanied by a labeled and numerically demarcated set of axes. The examinee's task is to plot the necessary points and use the appropriate tool (CURVE or LINE) to connect the points. Because the lines and curves are drawn by the appropriate tool, the examinee can focus more on the substance of the problem than the mechanics of responding.

Insert Figure 3 about here

The GM interface gives the examinee reasonably powerful editing capabilities. The examinee can erase points one at a time in the reverse order from which they were entered. He or she can also select a curve or line and delete a chosen point, after which the curve or line is immediately redrawn. Finally, an already-completed line or curve can be reshaped dynamically, simply by selecting a point and dragging it to a new location.

We have collected data about the GM interface from 20 students, most of whom were already enrolled in the first year of graduate school. All of these students reported that they typically wrote school papers on computer, and 19 of 20 indicated that they used a computer at least once a week. Most found the GM interface easy to use, though six stated that they encountered difficulty.

Much of the reported difficulty appeared to come from trying to enter points between grid lines; the interface snaps such points to the nearest intersection. This design decision was made so that the examinee could tell exactly where each point was placed. Any line or curve will obviously have intermediate points and, if the given scale is not used as a problem solving constraint, the examinee may generate intermediate points only to find that they cannot be plotted. One resolution is to give a default scale that will work for the item but, at the same time, allow redefinition so that the examinee can move intermediate points to the intersections as desired.

Automatic scoring. The automatic scoring for GM is a subset of the GE methodology. Thus, the key encodes the conditions that any response must meet to be creditable and scores responses by evaluating them against these conditions.

Consider the following question:

On the grid below, draw a triangle that has an area equal to 6 units.

The scoring key would essentially test the response to see if it:

- Contained three points,
- Had lines connecting the points, and
- Had an area equal to 6.

Though these conditions are clear and concise, specifying them precisely in computer code so that responses can be scored unambiguously requires effort and technical skill. For the GM response type to work effectively in high-volume, high-stakes programs, key creation must be efficient, undemanding of technical skill, and virtually fool-proof. These same conditions also hold for GE scoring.

To address this problem, we are using general item classes each of which has a single parameterized key. In principle, one would identify a relatively small number of very large general classes, and then create, test, and seal the parameterized keys. Test developers would use a simple tool to specify the parameters and test the key for any given item. The evaluation approach to ME already, in fact, works in much this way, where ME is the general class, the correct expression is the parameter, and the parameterized evaluation code is the single key that processes all responses in the class.

With respect to GM, an example of a general class (not yet implemented) might be questions requiring the construction of geometric objects with given dimensional characteristics. The sealed, uninitialized key would contain the code to determine whether the examinee's construction was of some to-be-specified type and dimension. Once the item had been written, the test developer would parameterize the key to reflect the conditions stated in the item stem, probably by clicking on a menu of options. Parameters would include the object (e.g., line, triangle, square, rectangle), dimensional characteristic of interest (e.g., length, width, slope, area, perimeter), one or more operators (e.g., equal to, greater than, less than), and one or more values. Note that the test developer interface would need to restrict the combination of some parameter types (e.g., line and area). Note also the variety of GM questions that could be scored using the elements given in this limited example.

Extended Constructed-Response Problems

We have argued that the three response types presented above can broaden, if only modestly, the conception of mathematical problem solving presented in computerized-adaptive tests. Even so, the tasks we used to illustrate these response types have been relatively short, elemental, and isolated. Such tasks arguably represent only a segment--and likely a small one at that--of the important problems encountered in higher education.

Consequently, it may be profitable to consider how we might create more extended problem types for use in computerized tests.

One possibility is to build sets around common stimuli that give the examinee a chance to become more immersed in a particular problem context--and the questions related to it--than can occur with a series of unrelated items. Figure 4 gives such a set adapted from the NCTM Standards (NCTM, 1989, p. 207), in which all three response types are combined (though not yet in computer-deliverable form). Within the same problem context, then, the examinee must deal with determined and under-determined tasks, and with quantitative and qualitative representations.

Insert Figure 4 about here

For computerized adaptive tests, sets pose challenges. For one, there may be some amount of conditional dependency among items that is not handled within the usual item response models. Models to account for such dependency are, however, emerging (e.g., Bradlow, Wainer, & Wang, in press). Also, there are inevitably issues of how to balance validity, efficiency, and generalizability. For example, if lengthier problems better represent the criterion setting, an acceptable level of generalizability may require more sets than available testing time would permit. Existing adaptive tests utilize sets in reading comprehension for similar validity goals but limit the number administered and intersperse other more efficient item types to maximize generalizability.

Conclusion

This paper presented three automatically scorable response types for computerized-adaptive admissions tests. We believe these response types have the potential to broaden the conception of mathematical problem solving embodied in such tests. This broadening involves using well-structured quantitative problems that are more open-ended, admitting under-determined problems, and giving greater emphasis to qualitative reasoning in both well-determined and more loosely structured situations.

Although we have focused on adaptive testing, these response types could also be used in other computerized testing frameworks like mastery or linear testing. Certainly, there is good precedent for employing extended constructed-response tasks in this way, as is the case with essays in the computerized Graduate Management Admission Test or with design simulations in the Architect Registration Examination (Bejar, 1995).

The work we've described is clearly in its beginning stages. One worthy objective is to connect more fully with current conceptions of mathematical problem solving. At a high level, it would seem desirable to create test designs that better reflect cognitive psychological accounts (e.g., Marshall, 1995), as well as discipline-based perspectives (e.g., NCTM, 1989) of mathematical proficiency. We would then use those designs to direct development of additional response types and their assembly into more theoretically coherent admissions test models. At a lower level, cognitive task analysis would be helpful in establishing that the resulting response

types did indeed call upon the processes implicated in these problem-solving conceptions.

A stronger theoretical connection would likely suggest extending the kinds of qualitative reasoning questions we can present. Informal reasoning in mathematics not only involves expression through graphs but also through diagrams, natural language, and other non-quantitative forms. We should be able to construct interfaces that present questions and allow examinees to manipulate objects like Venn diagrams to model underlying problem situations. Similarly, we should be able to permit examinees to construct short verbal explanations that justify their answers or that explain the problem situation. By constraining the language that can be used and employing advances in natural language processing, accurate scoring should be possible. Having multiple ways to represent problem situations opens up the possibility of presenting questions that require translating from one representational mode to another, or even of allowing examinees to choose the reasoning mode they are most comfortable with.

We and others have argued that the decision to use a new response type in place of conventional multiple-choice items needs to weigh a broadly defined view of benefits against costs (Bennett, 1993). For our three response types, we have as yet only limited data to inform this view. We have already suggested the benefit of improved construct representation and made a preliminary case largely on theoretical grounds. Among other things, we need to know more about how these response types differ from multiple choice in their psychometric functioning and in the solution processes examinees invoke when interacting with them. We need to know more about their fairness for women and minority group members. Also, we must understand something about their actual influences on teacher and student behavior (although we would expect a positive outcome by virtue of better construct coverage). Finally, we must complete analyses of scoring accuracy and of the costs of using these response types in operational settings.

Computer-based testing offers an unrealized flexibility that goes well beyond what we can feasibly do in large-scale paper-and-pencil admissions testing programs. Realizing that potential won't come immediately, easily, inexpensively, or without mistakes (and it may even require temporary retreats). Along with the more substantial innovations occurring in licensure testing (e.g., Bejar, 1995; Clauser et al., 1997), the modest attempts at innovation described in this paper hint at what may be possible in the not-too-distant future, both within and beyond the adaptive testing paradigm. We should expect that these innovations will accelerate as a critical mass of infrastructure, operational programs, and research effort coalesces, and as competition forces computer-based test providers to deliver ever greater value to their constituents.

References

Bejar, I. I. (1995). From adaptive testing to automated scoring of architectural simulations. In E. L. Mancall & P. G. Bashook (Ed.), Assessing clinical reasoning: The oral examination and alternative methods (pp. 115-130). Evanston, IL: American Board of Medical Specialties.

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. (1997). Speculations on the future of large-scale educational assessment (RR-97-14). Princeton, NJ: Educational Testing Service.

Bennett, R. E., & Bejar, I. I. (1997). Validity and automated scoring: It's not only the scoring (RR-97-13). Princeton, NJ: Educational Testing Service.

Bennett, R. E., Morley, M., Quardt, D., Singley, M. K., Katz, I. R., & Nhouyvanisvong, A. (in press). Psychometric and cognitive functioning of an under-determined computer-based response type for quantitative reasoning (RR-xx-xx). Princeton, NJ: Educational Testing Service.

Bennett, R. E., Steffen, M., Singley, M. K., Morley, M., & Jacquemin, D. (1997). Evaluating an automatically scorable, open-ended response type for measuring mathematical reasoning in computer-adaptive tests. Journal of Educational Measurement, 34, 163-177.

Bradlow, E. T., Wainer, H., & Wang, X. (in press). A Bayesian random effects model for testlets. Psychometrika.

Braswell, J., & Kupin, J. (1993). Item formats for assessment in mathematics. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement (pp 167-182). Hillsdale, NJ: Erlbaum.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. Journal of Educational Measurement, 29, 253-271.

Clauser, B. E., Ross, L. P., Clyman, S. G., Rose, K. M., Margolis, M. J., Nungester, R. J., Piemme, T. E., Chang, L., El-Bayoumi, G., Malakoff, G. L., & Pincetl, P. S. (1997). Development of a scoring algorithm to replace expert rating for scoring a complex performance-based assessment. Applied Measurement in Education, 10, 345-358.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. American Psychologist, 39, 193-202.

Gallagher, A., & Cahalan, C. (in progress). Examining the interface of mathematical expression items (RR-xx-xx). Princeton, NJ: Educational Testing Service.

Glaser, R. (1991). Expertise and assessment. In M. C. Wittrock & E. L. Baker (Eds.), Testing and cognition (pp. 17-30). Englewood Cliffs, NJ: Prentice-Hall.

Marshall, S. (1995). Schemas in problem solving. Cambridge, England: Cambridge University Press.

National Council of Teachers of Mathematics (NCTM). (1989). Curriculum and evaluation standards for school mathematics. Reston, VA: National Council of Teachers of Mathematics.

Nhouyvanisvong, A., Katz, I. R., & Singley, M. K. (1997, March). Toward a unified model of problem solving in well-determined and under-determined algebra word problems. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Simon, H. (1978). Information-processing theory of human problem solving. In W. K. Estes (Ed.), Handbook of learning and cognitive processes (pp. 271-295). Hillsdale, NJ: Erlbaum.

Table 1

A Broadened View of Mathematical Problem Solving in Terms of Structure and Representation

Problem Representation	Problem Structure: Well-structured	Problem Structure: Ill-structured
Quantitative	Questions in multiple-choice or open-end format that have a single best answer and emphasize manipulating mathematical forms.	Loosely formulated, open-ended questions that have multiple good answers and emphasize manipulating mathematical forms.
Qualitative	Questions in multiple-choice or open-ended format that have a single-best answer and emphasize manipulating graphical, diagrammatic, verbal or other non-quantitative forms.	Loosely formulated, open-ended questions that have multiple good answers and emphasize manipulating graphical, diagrammatic, verbal or other non-quantitative forms.

Table 2

Some Illustrative Mathematical Expression Items

1. If 12 eggs cost x cents and 20 slices of bacon cost y cents, what is the cost in cents of 2 eggs and 4 slices of bacon?
2. If n is the average (arithmetic mean) of the three numbers 6, 9, and k, what is the value of k in terms of n?
3. If a certain object has been moving at the constant rate of x meters per minute, how many meters has the object moved in the last y seconds? Express your answer in terms of x and y.

Note. Item #2 and #3 are adapted from GRE Practicing to Take the General Test Big Book, Princeton, NJ: Educational Testing Service, 1995.

Table 3

Some Illustrative Generating Examples Items

1. Joe is driving cross country. He travels 3,000 miles in 60 hours, switching cars somewhere along the way. The two cars have different average speeds, each of which does not exceed 70 miles per hour. Give an example of a speed and time for each leg of the trip.
2. If n and m are positive integers and $11n-7m=1$, what are two different possible sets of values for n and m?
3. A company makes a profit of $3.30 on every hardback book it sells and a profit of $1.20 on every paperback book it sells. Last month the company sold more than twice as many paperback books as hardback books and it made a profit of $3,960 on the books. How many books could the company have sold last month?
4. Two lines in the xy-coordinate plane are perpendicular. If the equation of the first line is $x + 5y = 17$, what is a possible equation, in the form $y = f(x)$, of the second line?

Table 4

Some Illustrative Graphical Modeling Items

1. Suppose you are driving at a constant speed from New York to Washington, DC, about 200 miles away. About 80 miles from New York you pass through Philadelphia, Pennsylvania. Sketch a graph of your distance from Philadelphia as a function of time from the beginning to the end of the trip. (x = time in hours, y = distance in miles)

2. In a certain state, the income tax is calculated as shown below:

Income	Tax
Less than \$10,000	0%
\$10,000 to \$20,000	3% of the income over \$10,000
Over \$20,000	\$300 plus 5% of the income over \$20,000

On the graph below, plot the relationship between the tax in dollars to income from \$0 to \$50,000. (x = income in dollars, y = tax in dollars)

3. Two years ago, John was taller than Susie. Last year, John and Susie were the same height. This year, Susie is taller than John. Plot possible growth curves for John and Susie for the past two years, assuming that each child has grown at least one inch per year. (x = time in months, y = height in inches)

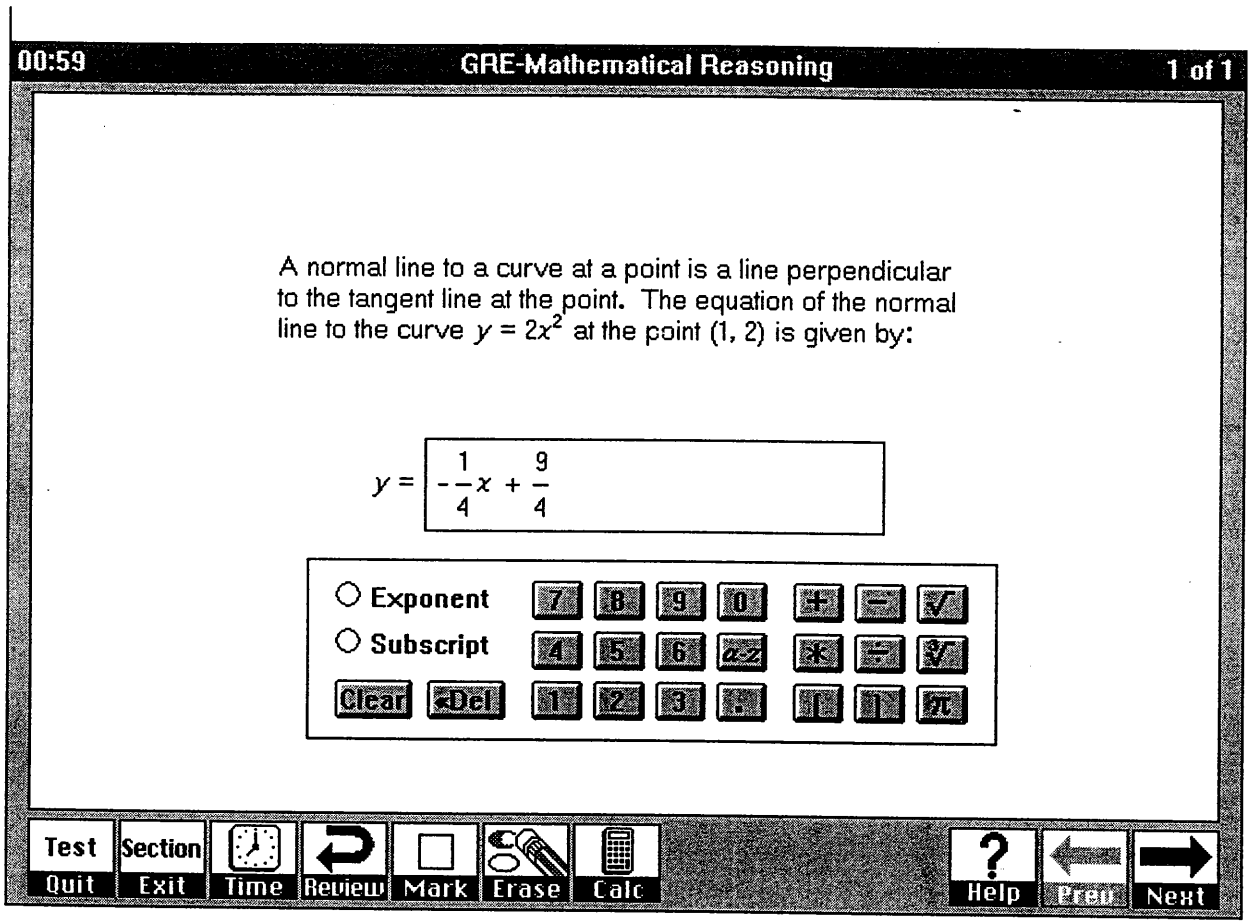
4. On the grid below, draw a triangle that has an area equal to 6 units.

5. Sketch a possible graph for a function that is decreasing everywhere, concave up for negative x and concave down for positive x. (x, y)

Note. Axis labels are in parentheses.

Figure 1

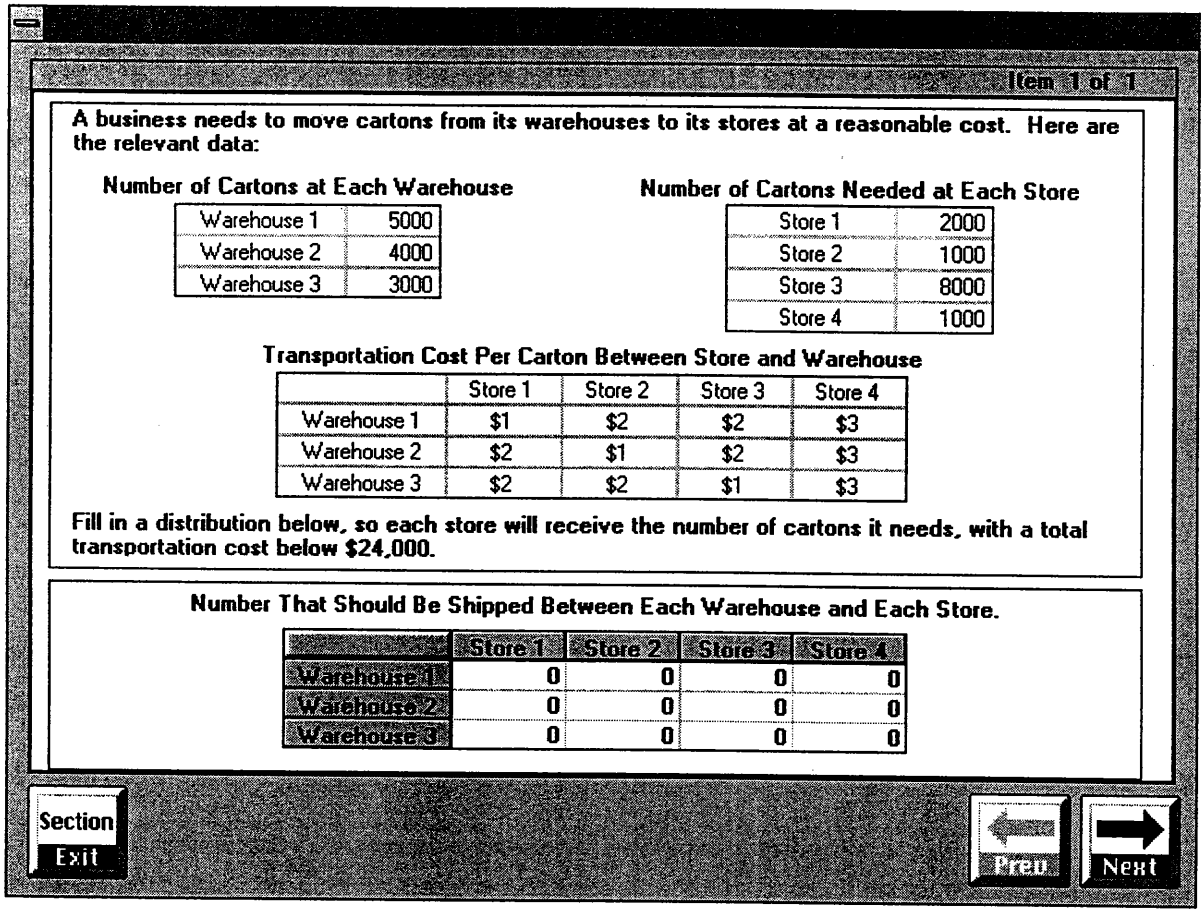
The Mathematical Expressions Interface with an Example Item and Correct Response



Note. Copyright (c) Educational Testing Service, 1996.

Figure 2

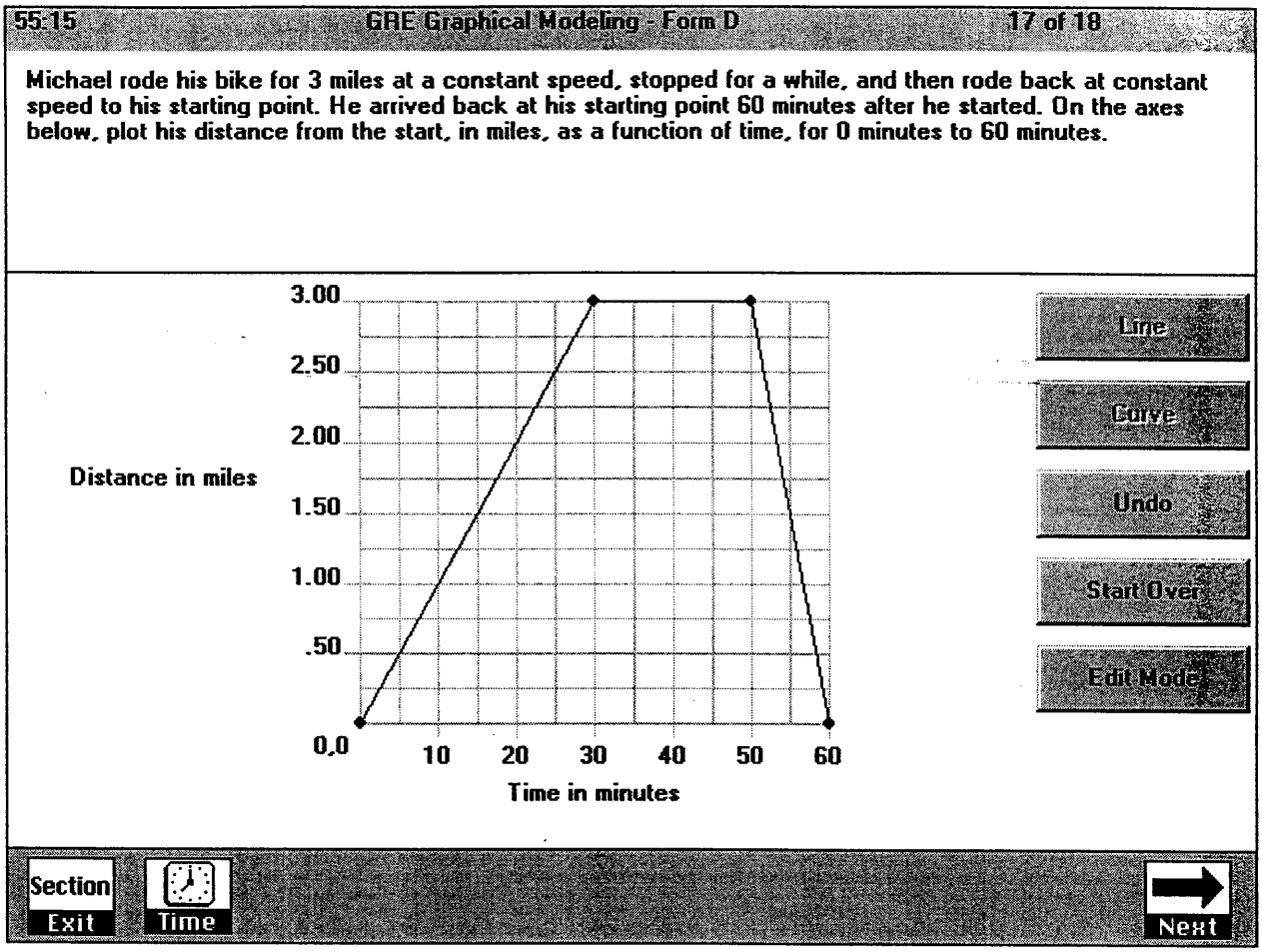
A Generating Examples Problem with its Interface



Note. Copyright (c) Educational Testing Service, 1997.

Figure 3

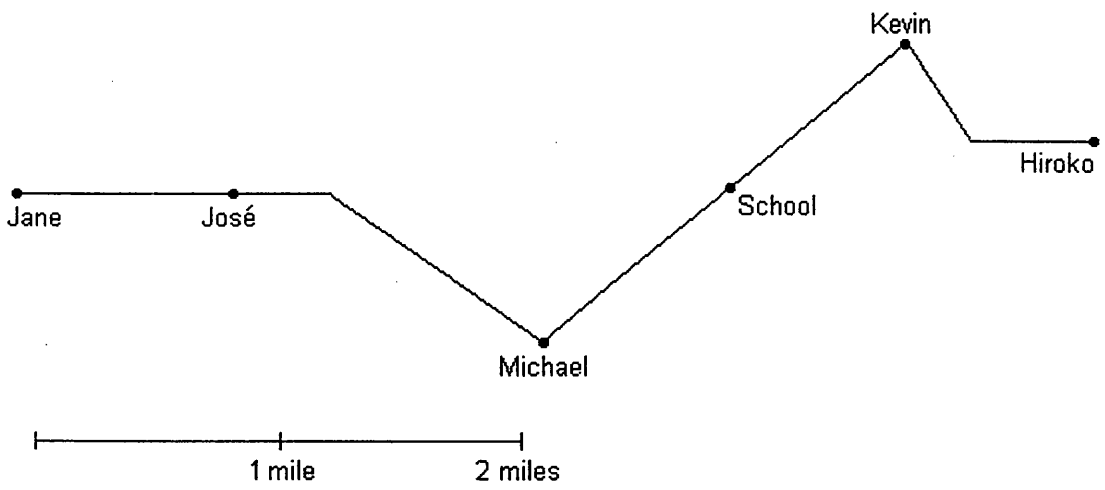
The Graphical Modeling Interface with a Sample Problem and Correct Answer



Note. Copyright (c) Educational Testing Service, 1998.

Figure 4

An Extended Constructed-Response Problem
Incorporating All Three Response Types



Jane, Jose, Michael, Kevin and Hiroko all travel along the same road on their way to school each morning. Kevin and Michael walk, Hiroko is driven by her father, and Jane and Jose ride bicycles to school. The map above shows the school and where each person lives.

1. Kevin walks to school at a speed of 4 miles per hour. Write an expression for the distance he is from school as a function of time, starting when he leaves his house.

2. Jane must get to school before 9:00 am but not before 8:30 am. She can ride her bike at any speed between 4 and 9 miles per hour, inclusive. Give a possible time she can leave and the speed at which she should ride her bike.

3. Hiroko's father is able to drive at 35 miles per hour along the straight sections of the road but has to slow down for the corners. Sketch a graph on the axes below to show how the car's speed varies along the route.

 Adapted from " Curriculum and evaluation standards for school mathematics" by the National Council of Teachers of Mathematics (NCTM), 1989, Reston, VA: National Council of Teachers of Mathematics.