The Alberta Journal of Educational Research *Vol. XLV, No. 4, Winter 1999, 409-427*

Jacqueline P. Leighton
Yale University

W. Todd Rogers

and

Thomas O. Maguire
University of Alberta

Assessment of Student Problem-Solving on Ill-Defined Tasks

Investigations of formal problem-solving are conducted with the expectation that they will predict or at least help understand informal or everyday problem-solving. For instance, if a student scores well on a multiple-choice physics exam, the expectation is that the student will also do well on an everyday physics problem. Traditionally the evaluation of problem-solving skills in educational testing and cognitive psychology has been dominated by formal, objectively scored tests, for example, multiple-choice tests (Garnham & Oakhill, 1994; Hambleton & Murphy, 1992). The relationship between formal and informal processes is questionable, however (Galotti, 1989). Formal tests may not elicit the same cognitive processes as informal tasks because they lack the process authenticity of informal tasks (Royer, Cisero, & Carlo, 1993). To address the lack of process authenticity, problem-solving skills can be directly evaluated using tasks that are "ill defined" and therefore more likely to elicit the cognitive processes associated with informal, everyday tasks. The purpose of the present study was to construct informal, performance tasks to evaluate both junior and senior high school students' problem-solving in mathematics. The task for students was to evaluate other students' solutions to two questions in mathematics. Results indicate that higher-achieving students generally preferred responses reflecting multiple approaches to problem-solving. A smaller number of students were also interviewed individually and asked to think aloud as they evaluated the solutions. Results indicate that students found multiple approaches to problem-solving desirable, while at the same time exhibiting problem-solving biases.

Les études sur la résolution formelle de problèmes sont entreprises dans l'attente qu'elles prédiront, ou du moins aideront à comprendre, la résolution quotidienne ou informelle de problèmes. Par exemple, si un élève réussit bien à un examen de physique à choix multiples, on s'attend à ce que sa performance soit aussi positive lors de la résolution d'un problème quotidien concernant la physique. Traditionnellement, l'évaluation des habiletés de résolution de problèmes dans les tests de rendement et dans la psychologie cognitive a été dominée par des examens formels et objectifs tels ceux à choix multiples (Garnham & Oakhill, 1994; Hambleton & Murphy, 1992). Cependant, le rapport entre les processus formel et informel est discutable (Galotti, 1989). Les examens formels, n'ayant pas l'authenticité procédurale des tâches informelles, pourraient ne pas faire appel aux mêmes processus cognitifs que les examens informels (Royer, Cisero, & Carlo, 1993). Pour parer au manque d'authenticité

Jacqueline Leighton is a Social Sciences and Humanities Research Council postdoctoral fellow in the Department of Psychology. She specializes in reasoning and problem-solving.
Todd Rogers is a professor in the Department of Educational Psychology and Director of the Centre for Research in Applied Measurement and Evaluation. He specializes in measurement and evaluation.
Tom Maguire is professor emeritus in the Department of Educational Psychology. He specializes in measurement and evaluation.

procédurale, on peut évaluer les habiletés de résolution de problèmes en employant des tâches qui sont "mal définies" et donc plus aptes à faire appel aux processus cognitifs associés aux tâches informelles de tous les jours. Le but de la présente étude était de créer des tâches informelles pour évaluer la capacité de résolution de problèmes en mathématiques d'élèves de la 7^e *à la 12*^{e} *année. La tâche des élèves était d'évaluer les solutions que les autres élèves avaient fournies à deux questions de mathématiques. Les résultats indiquent qu'en général, les élèves les plus performants préfèrent les réponses impliquant des approches multiples à la résolution de problèmes. Lors d'entrevues individuelles avec un plus petit nombre d'élèves, on leur a demandé de réfléchir à haut voix en évaluant les solutions. Les résultats indiquent que les élèves jugent désirables les approches multiples à la résolution de problèmes, tout en démontrant des préjugés quant au processus de résolution.*

Schools are experiencing a shift toward more performance-based assessment, that is, assessment that measures problem-solving skills as manifested in everyday tasks (Royer, Cisero, & Carlo, 1993). Wiley (1991) defines everyday tasks as *learning tasks* when they mirror instructional activities or *life tasks* when they mirror real-life activities (for which learning tasks prepare the student). This move is an important one because of the increasing concern that traditional formal tasks used in educational assessment, such as multiple-choice and true-false tests, do not measure the same cognitive strategies or processes as do informal or performance-based assessments (Galotti, 1989; Wiley, 1991). Although the terms *formal* and *informal* are not normally used to describe types of educational assessment, these terms are commonly used in the cognitive psychological literature to describe problem-solving tasks (Garnham & Oakhill, 1994). As in education, the field of problem-solving in psychology is also experiencing a shift from traditional formal tasks, such as conditional and categorical syllogisms, to informal or performance-based assessment (Cummins, 1995). This shift represents a desire on the part of educators and psychologists to increase the validity of both problem-solving tasks and the inferences made from them about students.

The shift is taking place because formal tasks are believed to lack *process authenticity* insofar as they fail to measure the same cognitive strategies as those measured by informal, performance-based tasks (Royer et al., 1993). In addition, formal tasks are believed to lack *task authenticity* insofar as they do not represent the kind of meaningful and relevant tasks found in the real world (Royer et al., 1993; Rogers, Maguire, & Leighton, 1998). The use of tasks that have both task and process authenticity is important in order to "more accurately mirror and measure what we value in education [and psychology]" (Hart, 1994, p. 9).

Although many terms are used to describe alternatives to traditional, formal assessment (e.g., *authentic assessment, holistic assessment, outcome-based assessment,* and the term we use in this article—*performance-based assessment*), they all implicitly suggest a move toward assessment that more closely mirrors the goal of education—to prepare students to deal effectively with real-life problems. As Hart (1994) describes, "An assessment is authentic when it involves students in tasks that are worthwhile, significant, and meaningful" (p. 9). Such tasks include problems students might face in their everyday environment wherein they have to plan, organize, evaluate, and/or synthesize knowledge in order to generate a response instead of simply choosing the best option. Use of performance-based assessment requires students to use higher-order thinking strategies and a broad range of knowledge, just as people do when they need to solve everyday problems such as evaluating health care alternatives or the platforms associated with different political parties (Hart, 1994; Tombari & Borich, 1999). In return, use of performance-based assessment can inform students about where they need to expand their knowledge and/or strategies in relation to themselves instead of others (Tombari & Borich, 1999).

Formal versus Informal Tasks

Although there are many differences between formal and informal tasks, formal tasks are used in educational assessment because they are less problematic to employ than are informal tasks. To be sure, ease of implementation is an important variable to consider, especially in instructional settings, but it should not overshadow valid assessment.

A distinction is usually drawn between formal and informal tasks in educational assessment as well as in psychology (Hambleton & Murphy, 1992; Galotti, 1989; Garnham & Oakhill, 1994). This distinction is most clearly understood by highlighting the characteristics associated with formal and informal tasks. For instance, formal tasks in education and psychology are alike in that they (a) normally hold all relevant information leading to a solution, (b) are self-contained, (c) involve a single correct answer, (d) can be solved using conventional procedures, (e) involve solutions that are unambiguous, (f) entail topics that are typically narrow and only of academic interest, and (g) do not necessarily prepare students to solve other problems successfully (Galotti, 1989; Hambleton & Murphy, 1992). In contrast, informal tasks tend to have the opposite characteristics. In particular, students normally must search for solutions to informal tasks by considering additional information to that presented in the task. By considering outside information, students often will consider distinct procedures for generating solutions, and as a result more than a single correct solution to the task is likely to be extended (Garnham & Oakhill, 1994).

In contrast to formal tasks, informal tasks are infrequently used and, consequently, have attracted little empirical attention (Garnham & Oakhill, 1994). The frequent use of formal tasks to the detriment of informal tasks in both the educational and psychological spheres has a common explanation. On the one hand, research with formal tasks lends itself to classroom use and laboratory study because the tasks are self-contained, permit the instructor (or investigator) to control the information students use to solve the task (as the instructor provides the required and relevant information needed), and allow an unambiguous evaluation of students' performance (Galotti, 1989). Hence formal tasks have come to be associated with a well-defined methodology (Garnham & Oakhill, 1994). On the other hand, use of informal tasks is considered risky because performance-based assessments are not easily designed or administered in classroom and laboratory settings. For example, informal tasks normally fail to reach acceptable levels of reliability (Hambleton & Murphy, 1992; Linn, 1994). Furthermore, informal tasks make it almost impossible for the instructor to control the information students use to solve the task because, by definition, performance-based assessments require students to consider information outside the task. Finally, students' responses to informal tasks are not easily scored because multiple approaches to reach a solution and a variety of solutions are possible (Galotti, 1989; Perkins, 1986). In sum, the strengths of formal tasks have overshadowed those of informal tasks (Hambleton & Murphy, 1992; Royer et al., 1993).

Formal tasks, however, are generally used with the implicit expectation that these tasks will predict performance on informal or everyday problem-solving tasks (Galotti, 1989; Harman, 1995). Although empirical results obtained from the use of formal tasks have spawned many conclusions about formal problem-solving processes, these conclusions have not been informative about everyday cognition or performance on everyday problem-solving tasks (Galotti, 1989; Hambleton & Murphy, 1992; Rogers et al., 1998). Research with informal tasks is therefore needed in order to understand everyday problem-solving (Perkins, 1986).

Real Tasks and Real Problem-Solving

Although students' problem-solving processes and products are typically assessed with formal tasks, the real aim of these measures is to predict process and product on informal or everyday tasks (Galotti, 1989; Rogers et al., 1998). In educational testing, for example, where achievement has traditionally been prescribed in terms of behavioral outcomes, the implicit aim is to predict cognitive processes and structures outside of the testing situation (Norris, 1983, 1985). To be sure, this is the main goal of education—to prepare students to deal effectively with real-life problems whether they come in the form of balancing a checkbook or evaluating the platforms associated with different political parties. If the goal is to make inferences about process and product on everyday, real-life tasks, then the question of why investigators employ formal tasks, which may not measure the cognitive processes elicited in everyday tasks, is worth considering. As mentioned above, formal tasks have come to be associated with a stronger methodology. A strong methodology is especially important in educational testing where high-stakes decisions about students are made from achievement results (Hart, 1994).

Although some empirical attention has been devoted to studying tasks that incorporate real-life problems and measure the kinds of processes that individuals use in everyday contexts, additional research is required (Maguire, Hattie, & Haig, 1993; Royer et al., 1993; Wiley, 1991). In particular, additional research is needed (a) to identify the processes students engage when they solve everyday performance-based assessments (Sugrue, 1995) and (b) to devise methods by which performance-based assessments can be efficiently scored. This call is also being heeded in the psychological literature (Evans, Over, & Manktelow, 1993; Galotti, 1989; Gigerenzer & Murray, 1987). If the transition from formal, objective assessment to informal, performance-based assessment is to be made, then it must be done with proof of rigor and efficiency. The goals of this article are to describe the development of two performance-based, mathematics items that can be "objectively" scored and to present the results obtained when they were administered to students in grades 9 and 10.

Math Items: Pack the Pop and Attitude and L.A. Scores

The administration of both "Pack the Pop" and "Attitude and L.A. Scores" followed a similar procedure. The two performance-based assessments were developed so that they would match the characteristics of informal tasks; that is, they were designed to (a) be relevant and interesting, (b) require students to consider multiple problem-solving approaches (and solutions), and (c) invoke higher-level thinking strategies in students such as planning and evaluating. Keeping with the goal of efficiency, the items were designed to be group-administered. Students were required to read the problem and then to evaluate other students' solutions of the problem. In this respect, students engaged in everyday problem-solving because they considered an ill-defined problem and meta-evaluated or evaluated how others solved the problem.

Participants

The students who participated in this study were enrolled in grades 9 and 10 at a junior and senior high school respectively, located in Edmonton, Alberta. One hundred and seventy students participated, 89 (32 boys and 57 girls) in grade 9, and 66 (31 boys and 35 girls) in grade 10. The items were administered in class as well as individually. In particular, groups of 15 to 20 students completed the items in their classes. Interviews were conducted with four grade 9 students and eight grade 10 students.[1] For individual interviews, teachers selected students who were generally highly verbal and able to think aloud as they solved the item.

Item 1: Pack the Pop

Pack the Pop was developed as follows: First, using a think-aloud protocol (Ericsson & Simon, 1993), the problem was administered individually to two small samples of students (n=5; n=6) in order to obtain their impressions, possible solutions, and a record of how they obtained their solutions. Second, their responses and explanations were used to modify the problem and construct the "student solutions" that were presented to the students in the present study for evaluation. The *final* statement of the problem read as follows:

> Suppose that you are going to have a party and that you have decided to ride your bike to the store to buy some cans of pop. On the back of your bike is a carrier with a box in which you will carry the cans. This box is 32 cm long, 19 cm wide, and 27 cm high. It has a flat lid. Each can of pop is 12.5 cm high, has a 6 cm diameter, and has a volume of 355 mL. If the cans are stacked in one direction (all vertically or all horizontally) on top of one another, what is the maximum number of cans you can carry in the box?

The problem was followed by an outline of the evaluation procedure to be used:

> Presented on the next few pages are the steps used by 4 different students as they determined the maximum number of pop cans that would fit in the box on the back of the bike. Pretend that you are their teacher, and you are marking their answers.
>
> Each answer is divided into three parts:[2] *a starting idea, a basic procedure*, and an *answer*. After each part, you will be asked to indicate what you think about what the student did using the following scale. Use the scale this way: If you think that the student's starting idea is very good, then you would circle the *very good* option (a number "5" on the scale). If you think that the student's starting idea is okay, then you would circle the *good* option (a number "4" on the scale). If you think that the student's starting idea is bad, then you would circle the *bad* option

> (a number "2" on the scale). If you think that the student's starting idea is very bad, then you would circle the *very bad* option (a number "1" on the scale). If you really do not know whether the student's starting idea is good or bad, then you would circle the *don't know* option (a number "3" on the scale). Use this same marking procedure when you mark the student's basic procedure and the student's answer as well.

During group as well as individual administrations, students were given as much time as they needed to evaluate the four answers. With individual interviews, students were asked to think aloud as they completed the item. To this end, standard probes were used to clarify students' thoughts such as "What did you mean when you said ...?" "What were you doing there?" and "Is there something else you wish to add before we go on to the next student?" (Ericsson & Simon, 1993). Finally, students were not explicitly trained to use the evaluation scheme precisely because our primary interest was to assess students' own meta-evaluative strategies and not whether they could be trained to evaluate according to our intentions.

The first solution students were asked to evaluate involved stacking all the cans vertically, which results in 30 cans fitting inside the box. The second solution, algorithmic in nature, involved using the equation for volume (i.e., v=lwh) to find the volume of the box and then dividing this value by the volume of the can; the result is 46 cans. The third solution involved stacking all the cans horizontally, which results in 24 cans fitting in the box. Finally, the last solution involved stacking the cans first horizontally and then vertically to find out which method yields the greater number of cans that fit inside the box.

Results[3]

We decided the "ideal" ratings for the four solutions after careful consideration of each solution. It was decided that *Solution 1* should be rated a 4 or 5 (i.e., good or very good) for its starting point because it starts by stating the known and relevant information to focus the immediate problem. Please note that after some consideration we decided not to differentiate between the scale points 4 and 5 because students appeared to have used these ratings similarly depending on how conservatively they interpreted the scale. For similar reasons, scale points 1 and 2 (i.e., very bad and bad respectively) were also combined. Both ratings at the top or bottom indicate a generally positive or negative evaluation respectively. Continuing with solution 1's evaluation, its basic procedure should be rated a 3 (i.e., don't know) because it fails to suggest alternate procedures for stacking the cans, and its final answer should be rated a 2 or 1 because it fails to mention alternate solutions to the problem (i.e., stacking the cans in a different direction).

Solution 2 should be rated a 2 for its starting point because the question asks for the maximum number of cans that fit in the box and not for how much liquid soda can be poured in the box; a 1 or 2 for its basic procedure because the response continues to ignore the physical component of the cans; and 1 or 2 for its final answer because the solution continues to neglect the physical component of the cans. Solution 3 should be rated similarly to solution 1 because it is analogous to solution 1 with the exception that the cans are stacked horizontally instead of vertically. Finally, solution 4 should be rated a 4 or 5 for its starting point because it states the relevant information and mentions at least two ways of solving the problem; a 4 or 5 for its basic procedure because it demonstrates how to solve the problem in at least two ways; and 4 or 5 for its final answer because it recommends the procedure that yields the maximum number of cans—stacking the cans vertically.

Class administration. The distributions of ratings assigned by grade 9 and grade 10 students to the item Pack the Pop are shown in Tables 1 and 2 respectively. Only ratings of the final answer to each solution are shown because these ratings, more so than ratings of the starting point and basic procedure, distinguished students at different math achievement levels (termed *quartiles* in Tables 1 and 2 and explained below). We believe the ratings of the final answer are especially informative because this is the point at which students can combine previous information from the starting point and basic procedure together with the final answer to rate the overall solution. Unlike final answer ratings, the ratings provided for the starting point and basic procedure are less informative because students rate these sections before they have all available information about the solution. We consider the ratings assigned to the final answer of each solution to represent the best available evidence of students' overall impression of the solution.

Notice in Tables 1 and 2 that students are divided into quartiles, an index of mathematics achievement. This index was created using the math grades, expressed as percentages, from the third (April) report cards. Math grades on each class list were divided into four categories, abiding by clear cut-points whenever possible. Categories were created anew in each class because different teachers used different criteria for assigning grades.[4] Typically, students in the first quartile received grades below 40%, students in the second quartile achieved grades between 41 and 59%, students in the third quartile achieved grades between 60 and 79%, and students in the fourth quartile achieved grades above 80%. Although this index is fairly coarse, it provides a useful indicator of how students at each achievement level evaluated each of the four solutions.

Table 1
Percentage of Grade 9 Students' Assigned Ratings to Pack the Pop's Final Answer by Quartile

	Quartile 1[a]				*Quartile 2*				*Quartile 3*				*Quartile 4*			
	G[b]	*D*	*B*	$\bar{x}$[c]	*G*	*D*	*B*	$\bar{x}$	*G*	*D*	*B*	$\bar{x}$	*G*	*D*	*B*	$\bar{x}$
Solution 1	47	21	32	3.2	55	18	27	3.4	61	18	18	3.6	82	9	9	4.0
Solution 2	58	21	21	3.6	73	18	9	4.0	54	0	46	3.0	27	9	64	2.5
Solution 3	37	21	42	2.8	55	18	27	3.2	32	21	46	2.8	41	14	45	2.9
Solution 4	37	32	26	3.2	77	14	9	4.1	71	18	11	4.0	86	5	9	4.1

Note. Percentages may not add to 100 either due to rounding or missing cases; standard deviations for average scale ratings center around 1.0, with a range of .6-1.4.
[a]Quartile 1, *n*=19; Quartile 2, *n*=22, Quartile 3, *n*=28, and Quartile 4, *n*=22.
[b]Scale ratings where G=very good/good, D=don't know, and B=very bad/bad.
[c]Average scale rating of final answer by quartile.

Table 2
Percentage of Grade 10 Students' Assigned Ratings to Pack the Pop's Final Answer by Quartile

	Quartile 1[a]				*Quartile 2*				*Quartile 3*				*Quartile 4*			
	G[b]	*D*	*B*	$\bar{x}$[c]	*G*	*D*	*B*	$\bar{x}$	*G*	*D*	*B*	$\bar{x}$	*G*	*D*	*B*	$\bar{x}$
Solution 1	75	13	13	3.9	74	11	16	3.7	76	6	18	3.8	83	0	8	4.2
Solution 2	19	6	75	2.3	32	26	42	3.1	41	24	35	3.1	0	50	50	2.3
Solution 3	25	19	56	2.6	42	26	26	3.1	47	29	24	3.3	58	33	8	3.4
Solution 4	75	6	19	4.0	79	5	11	4.2	82	12	6	4.2	92	0	8	4.2

Note. Percentages may not add to 100 either due to rounding or missing cases; standard deviations for average scale ratings center around 1.0, with a range of .5-1.4.
[a] Quartile 1, *n*=16; Quartile 2, *n*=19, Quartile 3, *n*=17, and Quartile 4, *n*=12.
[b] Scale ratings where G=very good/good, D=don't know, and B=very bad/bad.
[c] Average scale rating of final answer by quartile.

The Kruskal-Wallis test was used to examine how students at different quartiles rated the final answers of each solution. This test was used because we wished to be conservative about the assumptions we made about the data. Because responses to this item have not been previously collected, we were unjustified in assuming a normal distribution of ratings and in assuming knowledge of associated population parameters (Marascuilo & Serlin, 1988). Using the Kruskal-Wallis test, we found a significant main effect of grade for ratings of solution 2's final answer (*U*=3663.5, *df*=1, *p*=.0288). In particular, students in grade 9 rated solution 2's final answer more highly than did students in grade 10. Fifty-three percent of grade 9 students rated solution 2's final answer as *very good or good*, whereas at grade 10 only 23% of students rated this solution similarly. We also found a significant main effect of quartile for ratings of solution 1's and solution 2's final answers (solution 1: *H*=7.8276, *df*=3, *p*=.0497; solution 2: *H*=12.9087, *df*=3, *p*=.0048). Specifically, students at quartiles 3 and 4 rated solution 1's final answer more highly than did students in quartiles 1 and 2. Seventy-six percent of students at quartiles 3 and 4 rated solution 1's final answer as *very good or good*, whereas only 63% of students at quartiles 1 and 2 did the same. In addition, students at quartiles 1, 2, and 3 rated solution 2's final answer more highly than did students at quartile 4. Recall that solution 2 is considered to hold the poorest answer to the problem because it fails to take into consideration the physical properties of the cans. Forty-six percent of students at quartiles 1, 2, and 3 rated solution 2's final answer as *very good or good*, whereas only 14% of students at quartile 4 did the same.

We also looked at simple effects by testing whether there were any differences in ratings in specific grades and specific quartiles. For example, in grade 9 we found significant simple effects of quartile for ratings of solutions 2 and 4. Grade 9 students at quartiles 1, 2, and 3 rated solution 2's final answer more highly than did students at quartile 4 (*H*=14.2119, *df*=3, *p*=.0026). In contrast, grade 9 students at quartiles 2, 3, and 4 rated solution 4's final answer more highly than did students at the lowest quartile (*H*=10.2054, *df*=3, *p*=.0169). Seventy-eight percent of students at quartiles 2, 3, and 4 rated solution 4's final answer as *very good or good,* whereas only 37% of students at quartile 1 did the same. We did not find any significant simple effects in grade 10.

We also looked at differences in ratings in specific quartiles. For example, in quartile 1 we found significant simple effects of grade for ratings of solutions 2 and 4. For solution 2's final answer, grade 9 students rated this final answer more highly than did grade 10 students (*U*=243.5000, *df*=1, *p*=.0018). Fifty-eight percent of grade 9 students at quartile 1 rated solution 2's final answer as *very good or good,* whereas only 19% of grade 10 students at quartile 1 rated it as such. Conversely, grade 10 students rated solution 4's final answer more highly than did grade 9 students (*U*=84.5000, *df*=1, *p*=.0342). Seventy-five percent of grade 10 students at quartile 1 rated solution 4's final answer as *very good or good,* whereas only 37% of grade 9 students did the same. In quartile 2 we found significant simple effects of grade for ratings of solution 2's final answer. Grade 9 students rated solution 2's final answer more highly than did students in grade 10 (*U*=298.0000, *df*=1, *p*=.0164). Seventy-three percent of grade 9 students at quartile 2 rated solution 2's final answer as *very good or good,* whereas only 32% of grade 10 students rated it the same. We did not find any significant simple effects in quartiles 3 and 4.

In sum, we see that students in grade 10 are more critical of the quality of solution 2's final answer than are students in grade 9. In addition, students at higher quartiles tend to be more critical of solution 2's final answer than are students at lower quartiles, but more appreciative of solution 4's answer. We consider these results more closely in the Discussion section.

Individual administration. Although Tables 1 and 2 illustrate the ratings assigned to the final answers of each solution, these ratings do not illustrate the reasons for students rating the answers as they did. For this reason, Pack the Pop was administered individually to students so as to obtain their thoughts about particular solutions and ratings.

At the outset it is important to identify two characteristics that investigators of problem-solving have identified as important to sophisticated, formal reasoning. These characteristics include an appreciation of the multidimensionality of the problem and the selection of relevant information (Downing, Sternberg, & Ross, 1985; Evans, 1989; Evans et al., 1993; Johnson-Laird & Byrne, 1991; Pollard, 1990). The first of these involves an appreciation of possible alternative approaches to solving the task; that is, the acknowledgment that a single solution may not fully solve the problem and that others may exist. The second characteristic involves the ability to focus on relevant problem information within and outside of the task in order to reach a solution. Both these characteristics were observed in our interviews. For example, B.K., a grade 9 student at quartile 3, stated the following when asked why solution 1 was rated highest:

> because it was really clear ... everything was simple and you didn't have to remember lots of things you know. (1998)

Another grade 9 student, V.I., at quartile 4, who liked solution 4 more than any other solution said the following in support:

> I think is probably the best method because he determined the answer to the question. Like the question was what's the maximum number of cans you can fit

> in the box and he determined that it was 30. The others just came up with one step ... they didn't determine ... if like for this one [solution 3] he determined that 24 could fit into the box but he didn't determine that 30 could. He didn't do it vertically he did it horizontally. He didn't come up with two ways like this student did [solution 4] and I have to give this one [solution 4] the best. (1998)

Reflecting back on the previous solutions V.I. said the following:

> The others [solutions 1, 2, and 3] deserve lower marks because they didn't ... all of them ... they didn't determine ... like the last one [solution 4] determined both to see the maximum number of cans that could fit into the box ... but the others just came up with one solution but they don't know if that's the maximum number because they didn't figure out how many could fit vertically and how many could fit horizontally. Each of them [solutions 1, 2, and 3] came up with the right answers they just didn't answer the question completely. (1998)

A.C., a grade 10 student at quartile 3, gave the following reason for thinking solution 4 to be the best,

> Because he chose to do it two different methods. Like student 1 only did it one way and the other student did another way ... they all did one way and this student did it two ways which showed different perspectives of thinking of how it could go in and he is trying to find the best that he can do. (1998)

In particular, A.C. commented the following when evaluating solution 2:

> I think this one is wrong ... it could be wrong or right ... his answer is so much different compared to the other guy [solution 1] and I liked the logic of the other guy. (1998)

Only one grade 10 student, C.B., taking an advanced math curriculum (grade 11 math) pointed out why solution 2 was incorrect:

> She is finding the volume of the box but the thing is that they wanted stacked a certain way ... she has to find the length and stuff of the cans before she can find out how many can fit in because the volume might be a little bit more like 30.5 ... and that is not what they are asking ... This doesn't seem right ... she is finding how much pop it can hold and not the actual cans. The answer is wrong. Forty-six cans in the box? Of course there are going to be extra spaces because the cans aren't going to fit perfectly into the box. (1998)

In general, over 85% of the grade 9 and 10 students interviewed rated solution 4 highly and mentioned the sensibility of attempting at least two approaches when solving the problem; that is, they thought about the multidimensionality of the problem space. Also, they focused on relevant information more than on irrelevant information by considering the importance of attempting multiple approaches. Many of these students were at quartiles 3 and 4. Interestingly, some grade 9 students, but not grade 10 students, also rated solution 2 highly even though many mentioned that they were unsure about its procedure. Thus a point to consider is that grade 9 students may not be as sophisticated as grade 10 students in matching their ratings to their thoughts about a mathematical procedure under consideration.

Item 2: Attitude and L.A. Scores

Attitude and L.A. Scores was developed in the same fashion as Pack the Pop. As with Pack the Pop, item 2 requires students to evaluate other students' solutions, which were documented during the initial construction phases to the following mathematical problem.

> Mr. Smith, a Language Arts teacher, keeps telling his students about the importance of having a good attitude: Students with a positive attitude toward Language Arts will do better in Language Arts than students with a negative attitude. Now, he wants to show his students what the relationship between attitude and grades is, and how students' attitude scores can be used to predict grades.
>
> (a) First, Mr. Smith gives an attitude survey to his 30 grade 9 students. The attitude survey scores range from 0 to 100 points. A score near 0 means that the student has a very negative attitude toward Language Arts, and a score near 100 means that the student has a very positive attitude toward Language Arts.
>
> (b) One week later, Mr. Smith gives a Language Arts test. The Language Arts test scores range from 0 to 100 points as well. A score near 0 means that a student performed very poorly on the L.A. test, and a score near 100 means that a student performed extremely well on the L.A. test.
>
> (c) Then Mr. Smith draws the following graph where each dot corresponds to one student. *Please examine the graph (see Figure 1) below:*

After students read the problem, they were presented with the following task:

> Two questions about the information provided on the first page (i.e., math problem and graph) are presented on the following pages. DO NOT answer these questions yourself. Your task is to evaluate the responses of 6 students, much as a teacher would, to the two questions. The students' responses to each question are presented in steps. Each student's response is divided into three parts:
>
> (a) A starting point or a basic idea;
> (b) A procedure to find an answer; and
> (c) Final comment(s).

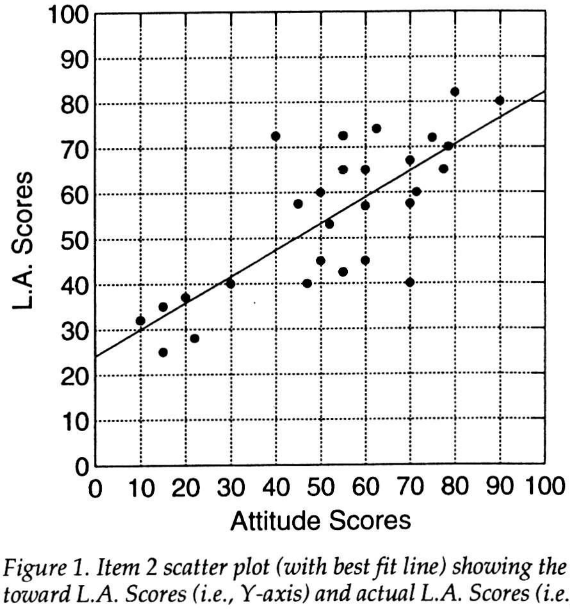

Figure 1. Item 2 scatter plot (with best fit line) showing the relationship between Attitude toward L.A. Scores (i.e., Y-axis) and actual L.A. Scores (i.e., X-axis).

> Question 1: John is a student in Mr. Smith's Language Arts class. He completed the attitude survey and he scored 55 on it, but he did NOT write the Language Arts test with the rest of the class because he was skipping class the day of the test. Mr. Smith needs to give everyone a score on the L.A. test. Mr. Smith asks you for help. What score do you think John would have obtained on the L.A. test, given that John scored 55 on the attitude survey?

The second question was identical to the first except that now a student named Susan missed the L.A. test because she had a doctor's appointment.

Students were instructed to evaluate the six solutions using the same scale and procedure used with Pack the Pop. After students read these instructions, they were given as much time as they needed to evaluate the six solutions to the two questions. The first solution involved finding John's attitude score on the scatter plot and deciding, from the three dots located at an attitude score of 55, the one that corresponded to his L.A. score. The second solution involved using the *line of best fit* to predict John's L.A. score. The third solution involved taking an average of the three dots located at an attitude score of 55 and using this new average value to predict John's L.A. score. The fourth solution involved using the general relationship between attitude and L.A. scores to predict John's L.A. score. The fifth solution involved using an emotional/psychological argument to predict John's L.A. score. Finally, the sixth solution involved using the general relationship between attitude and L.A. scores coupled with the observation that L.A. scores tend to be about 15 points higher than their corresponding attitude survey scores to predict John's score. The content of the six solutions to question 2 followed the same pattern. In addition, the students who were interviewed individually completed the item but were asked to think aloud as they evaluated each of the six solutions.

Results

After careful consideration of the solutions to item 2, we decided that the best rating of solutions were as follows: The starting point of solution 1 should be rated a 4/5 because it states the correct preliminary information, a 1/2 for its basic procedure because John's score is not represented on the graph as the solution claims it is (John did not take the L.A. test, so he does not have a coordinate pair on the graph), and a 1/2 for its final answer because it is based on incorrect information. The starting point, basic procedure, and final answer of solution 2 should be rated a 4/5, 4/5, and 4/5 respectively, because the best fit line is used to predict John's score; solution 2 is considered to be the best response to the problem. The starting point, basic procedure, and final answer of solution 3 should be rated a 4/5, 4/5, and 1/2 respectively, because John's score should not be chosen from any preexisting dots on the scatter plot as he is not represented by a dot on the scatter plot. The starting point, basic procedure, and final answer of solution 4 should be rated a 4/5, 4/5, and 1/2/3 respectively, because although this response correctly mentions the relationship between attitude scores and L.A. scores, it fails to use the best fit line to predict John's L.A. score. The starting point, basic procedure, and final answer of solution 5 should be rated a 1/2, 1/2, and 1/2 respectively, because this solution does not use any of the information presented in the scatter plot to predict John's score, but instead predicts John's score based on his tendencies to skip class. Finally,

Table 3
Percentage of Grade 9 Students' Assigned Ratings to the Final Answer of Both Question 1 and 2 of Attitude and Homework

					Question 1											
	Quartile 1[a]				Quartile 2				Quartile 3				Quartile 4			
	G[b]	D	B	$\bar{x}$[c]	G	D	B	$\bar{x}$	G	D	B	$\bar{x}$	G	D	B	$\bar{x}$
Solution 1	53	16	32	3.3	36	36	27	3.1	46	7	46	2.9	26	13	61	2.7
Solution 2	53	21	26	3.3	77	5	18	3.8	71	14	14	3.7	91	0	9	4.2
Solution 3	74	21	5	3.8	45	23	32	3.3	50	11	39	3.3	57	13	30	3.4
Solution 4	79	16	5	3.8	45	32	23	3.3	68	14	18	3.6	78	8	13	3.8
Solution 5	42	16	42	3.1	36	27	36	3.0	29	18	54	2.8	43	9	48	2.8
Solution 6	63	16	21	3.4	50	18	32	3.2	54	18	29	3.3	35	22	43	2.9
					Question 2											
	Quartile 1[a]				Quartile 2				Quartile 3				Quartile 4			
	G[b]	D	B	$\bar{x}$[c]	G	D	B	$\bar{x}$	G	D	B	$\bar{x}$	G	D	B	$\bar{x}$
Solution 1	53	26	21	3.4	48	5	48	2.8	54	11	36	3.1	39	26	35	3.1
Solution 2	58	21	21	3.4	62	14	24	3.6	61	11	29	3.4	87	9	4	4.0
Solution 3	53	16	32	3.2	62	19	19	3.4	63	19	19	3.6	52	26	22	3.4
Solution 4	42	32	26	3.2	38	29	33	3.0	32	11	57	2.7	35	17	48	2.7
Solution 5	68	16	16	3.6	38	14	48	2.9	25	25	50	2.7	26	30	43	2.6
Solution 6	26	21	53	2.7	30	30	40	2.9	18	29	54	2.6	9	18	73	2.1

Note. Percentages may not add to 100 either due to rounding or missing cases; standard deviations for average scale ratings center around 1.0, with a range of .7 - 1.4.
[a]Quartile 1, *n*=19; Quartile 2, *n*=22, Quartile 3, *n*=28, and Quartile 4, *n*=23.
[b]Scale ratings where G=very good/good, D=don't know, and B=very bad/bad.
[c]Average scale rating of final answer by quartile.

the starting point, basic procedure, and final answer of solution 6 should be rated 4/5, 3, and 1/2/3 because although this response notes the relationship between attitude and L.A. scores, it uses an idiosyncratic algorithm where a constant is added to individual attitude scores to predict corresponding L.A. scores. The same pattern of ratings follow for question 2 of this item.

Class administration. As with Pack the Pop, Tables 3 and 4 show students' ratings of final answers. As in Pack the Pop's analysis, students were divided into four quartiles, indexing mathematics achievement.

We again used the Kruskal-Wallis test to examine differences in ratings. For question 1 we did not find a significant main effect of grade, but there was a significant main effect of quartile for ratings of solution 2's final answer (H=7.951, df=1, p=.047). In particular, students at quartile 4 rated solution 2's final answer more highly than did students at quartiles 1, 2, and 3. Ninety-two percent of students at quartile 4 rated this final answer as *very good* or *good,* whereas only 73% of students at quartiles 1, 2, and 3 rated it similarly. Recall that we considered solution 2 to be the best response to question 1 because it involves using the best fit line to predict John's L.A. score. Finally, we did not find any significant simple effects in grade 9 or 10, but did find significant simple effects in quartile 2. In particular, in quartile 2, students in grade 10

Table 4
Percentage of Grade 10 Students' Assigned Ratings to the Final Answer of Both Question 1 and 2 of Attitude and Homework

	Question 1															
	Quartile 1[a]				*Quartile 2*				*Quartile 3*				*Quartile 4*			
	G[b]	*D*	*B*	$\bar{x}$[c]	*G*	*D*	*B*	$\bar{x}$	*G*	*D*	*B*	$\bar{x}$	*G*	*D*	*B*	$\bar{x}$
Solution 1	29	24	47	2.8	32	21	47	2.7	50	6	44	2.9	27	0	73	2.5
Solution 2	76	6	18	3.8	79	21	0	4.1	81	6	13	3.9	93	7	0	4.3
Solution 3	59	18	24	3.4	63	26	11	3.6	69	13	19	3.6	60	20	20	3.7
Solution 4	71	18	12	3.7	89	5	5	4.0	56	13	31	3.4	80	13	7	4.0
Solution 5	41	24	35	3.1	26	32	42	2.7	50	13	38	3.1	33	20	47	2.7
Solution 6	59	6	29	3.4	37	21	42	3.0	44	13	44	3.1	27	27	47	2.9
	Question 2															
	Quartile 1[a]				*Quartile 2*				*Quartile 3*				*Quartile 4*			
	G[b]	*D*	*B*	$\bar{x}$[c]	*G*	*D*	*B*	$\bar{x}$	*G*	*D*	*B*	$\bar{x}$	*G*	*D*	*B*	$\bar{x}$
Solution 1	50	19	31	3.3	47	11	42	2.9	67	7	27	3.3	43	14	43	2.9
Solution 2	69	23	8	3.6	68	11	21	3.7	73	7	20	3.7	86	7	7	4.0
Solution 3	50	29	21	3.4	58	32	11	3.6	67	27	7	3.6	50	36	14	3.7
Solution 4	50	21	29	3.7	37	26	37	4.0	47	20	33	3.4	29	14	57	4.0
Solution 5	23	31	46	2.8	26	16	58	2.6	47	20	33	3.2	7	43	50	2.6
Solution 6	36	7	57	2.7	26	21	53	2.5	40	20	40	2.9	21	14	64	2.4

Note. Percentages may not add to 100 either due to rounding or missing cases; standard deviations for average scale ratings center around 1.0, with a range of .4-1.4.
[a]Quartile 1, *n*=16; Quartile 2, *n*=19, Quartile 3, *n*=16, and Quartile 4, *n*=15.
[b]Scale ratings where G=very good/good, D=don't know, and B=very bad/bad.
[c]Average scale rating of final answer by quartile.

rated solution 4's final answer more highly than students in grade 9 (U=124.5, df=3, p=.014). Eighty-nine percent of grade 10 students rated this final answer as *very good* or *good*, whereas only 45% of grade 9 students rated it similarly. Recall that solution 4 uses the relationship between attitude scores and L.A. marks to predict John's score, but not the best fit line. No further significant simple effects were found.

For question 2, we again did not find a significant main effect of grade. We did, however, find a significant main effect of quartile for ratings of solution 2's final answer (H=8.854, df=3, p=.031). In particular, students at quartile 4 tended to rate solution 2's final answer more highly than did students at quartiles 1, 2, and 3. Eighty-seven percent of students at quartile 4 rated solution 2's final answer as *very good* or *good*, whereas 65% of students at quartiles 1, 2, and 3 did the same. We also found significant simple effects. In particular, in grade 9, students at quartile 1 rated solution 5's final answer more highly than did students at higher quartiles (H=10.365, df=3, p=.016). Sixty-eight percent of students at quartile 1 rated this solution as *very good* or *good*, whereas roughly 30% of students at quartiles 2, 3, and 4 did the same. Recall that solution 5 involves using John's tendency to skip class to predict his L.A. score. Also in grade 9, students at quartiles 1 and 2 rated solution 6's final answer more highly than did students at quartiles 3 and 4 (H=8.184, df=3, p=.042). Twenty-eight percent of students at quartiles 1 and 2 rated solution 6's final answer as *very good* or *good*, whereas roughly 14% of students at quartiles 3 and 4 did the same. Recall that solution 6 involves using an idiosyncratic algorithm to predict John's L.A. score. There were no simple effects in grade 10. Finally, in quartile 1, students in grade 9 rated solution 5's final answer more highly than did students in grade 10 (U=184.0, df=1, p=.015). No other significant simple effects were found.

Individual administration. As with item 1, item 2 was individually administered to students so that their thoughts and solution steps could be recorded. A number of interesting observations were made. First, as with item 1, some students focused and used irrelevant information in their evaluation of solutions. For instance, some students mentioned psychological factors or John's (Susan's) character in their assessment of solutions. In spite of this tendency, many were not sidetracked from focusing on other, more relevant variables. Focusing on relevant variables, however, did not always guarantee choosing the best solution if the variables were not considered thoroughly. For example, E.R., a grade 10 student at quartile 3 correctly focused on solution 2's use of the best fit line, but ended up rating the basic procedure as *don't know* despite prompts to detail the reasons for the rating.

> With the average line he is probably going to be near it and the others ... but I am not too sure about this one. (1998)

E.R. then went on to rate solution 2's final answer as *good* because:

> If that's the average for that [L.A. score] then he probably got around there. (1998)

However, E.R. did not end up selecting solution 2 as the best solution to question 1, but instead selected solutions 4 and 6 for the following reasons:

> They [solutions 4 and 6] are actually using the facts and the relationships, they are not just using presumptions or what other people ... they do not presume things like John didn't study for the test because he skips class ... they are using the facts given. (1998)

Another grade 10 student, J.O., at quartile 4 gave a psychological and a mathematical reason for rating solution 2's starting point highly:

> Everything this student put here was accurate and that he won't be on the graph because he didn't actually write the test but he did do the attitude survey ... and also because he was skipping the class it also says more about his attitude toward the class. (1998)

J.O. went on to rate the final answer of solution 2 as *completely right* for the following reasons:

> Because he didn't write the test you can't be sure ... John could be a very bright student with just a bad attitude toward class, but there is no way to know so the best thing you can do is just use the average line. (1998)

Although J.O. accepted comments about John's potential dislike of L.A., it did not appear to influence the ratings assigned to the solutions. For example, when rating solution 5's starting point, J.O. acknowledged that John could be an average student, but disagreed with using this information as the basis for a solution. J.O. provides the following reasons:

> Because he did skip the class it indicates something ... but I don't really agree that because he [John] got 55 it means he's an average student ... the test could have been extremely easy depending on what the average of the class was ... in fact, it could have been really high or really low ... so I think there needs to be more information given to determine if he is an average student. (1998)

J.O. finally selected solution 2 as the best response to question 1 providing the following reasons for the selection:

> Everything they seem to do is accurate ... they don't jump to any assumptions. You have to assume a little bit but ... based on what John did do which was the attitude survey and you have to assume the graph was plotted correctly and that the best fit line was drawn right. The only thing you really could do is find out what the average student who got 55 on the attitude test would score and the only way to do that is look at the best fit line. And everything that they did led up to doing that. (1998)

Another grade 10 student, L.B., at quartile 3 provided one of the most complete reasons for selecting solution 2 as the best solution to question 1 by contrasting it with solution 5:

> It's precise ... it uses the graph as evidence, it doesn't generalize whereas a lot of the other students [solutions] have sort of swayed from the question ... they have generalized and they haven't been precise enough in the steps. And that one person [solution 5] was way too interpretive of the person [John] as opposed to the evidence given ... like the evidence about the graph ... didn't use that at all and just used the information that the person was skipping and therefore they didn't study and they scored badly. (1998)

In many of the above comments it is noticeable that students understand the unavoidable uncertainty associated with prediction. This is an important point to understand if the best fit line is to be considered a worthwhile approach, because its use will only lead to judgments about the likelihood of an event. For example, L.B. acknowledges this uncertainty in the following evaluation, which is representative of the comments obtained from many of the students interviewed, of solution 4's final answer:

> It [solution 4] says "I cannot be positive John would have scored a 60 (he could score above or below this number), but the trend of scores in the graph suggest that 60 is a likely score" ... and I think this is probably pretty true ... well it is [true] because the average ... you calculate the average from the other students. (1998)

Finally, roughly 75% of the students who accepted psychological explanations also rated solution 2 highest, that is, they recognized that using the best fit line was the best approach to use in predicting L.A. scores. This suggests that the process of reasoning does not need to have a straightforward relationship to its product because students entertained nonmathematical justifications when evaluating some of the solutions, but they ultimately rated the best mathematical solution highest. Alternatively, it suggests that some (older) students are much more sophisticated at separating mathematical solutions from possible nonmathematical justifications. To be sure, if one were to only look at these students' ratings, one might not have guessed that some irrelevant information played a role in their evaluation of solutions.

Discussion

From the results presented it is possible to draw a number of conclusions about informal, performance-based assessment. First, just as with formal, objectively scored tests, the development of challenging and interpretable items for students is arduous work, and perhaps more so because there are fewer established guidelines to follow. After having developed and administered objectively-scored performance-based assessments such as Pack the Pop and Attitude and L.A. Scores, results suggest that such assessments can begin to discriminate between low and high achievers. For example, students at quartile 4 generally rated the best solution (i.e., solution 4 to Pack the Pop and solution 2 to Attitude and Homework) more highly than did students at lower quartiles. In addition, students at quartile 1 generally rated the poorest solution (e.g., solution 2 to Pack the Pop and solution 5 to Attitude and Homework) more highly than did students at quartiles 2, 3, and 4. Moreover, our items also discriminated between students in different grades. In general, students in grade 10 rated the best solution to both Pack the Pop and Attitude and Homework more highly than did students in grade 9. Finally, these items can be quickly scored because students can indicate their evaluations using the rating scales.

These assessments also appear to elicit the cognitive processes used to solve real, everyday tasks. For example, Pack the Pop's interview reports indicate that students who rated solution 4 highest tended also to point out the desirability of considering multiple approaches to solve a problem. In contrast, students who rated solution 2 highly tended to focus on the quickness of its algorithmic approach without considering its implications. These results to a great extent mirror the problem-solving of everyday tasks insofar as multiplicity of thought and foresight can avert poor solutions to problems.

Moreover, Attitude and L.A. Scores interview reports indicate that students appear to use nonmathematical heuristics, such as focusing on the psychological attributes of a person, to aid their problem-solving. Such heuristics have been noted by problem-solving researchers as prevalent in everyday reasoning because they yield quick and often reliable solutions with a minimum of effort (Evans, 1989; Newell & Simon, 1972; Tversky & Kahneman, 1982). Interestingly, our results indicate that some students who employ such heuristics are also able to choose the best mathematical solution. This suggests that even with performance-based assessment, what you see is not always what you get; that is, students may choose the best mathematical answer, but still entertain nonmathematical heuristics to assess solutions and ultimately arrive at a final response. In sum, our results provide preliminary evidence that informal, performance tasks can (a) be efficiently administered and evaluated if cast in an objectively-scored format, and (b) discriminate between students of different achievement levels.

Finally, a number of interesting points may be made about students' oral reports. At the outset, it was found by all interviewers that students in general had difficulty articulating ideas and/or thoughts about mathematical tasks (Kulak & Rooney, 1999; Robinson, 1999). Although this may prove to be a limitation to the validation of performance items in mathematics, recent research suggests that retrospective reports may be more accurate than concurrent reports (Robinson, 1999). In the future, special attention may be directed at students' ability to articulate mathematical ideas in the classroom. Furthermore, students exhibited characteristics that cognitive psychologists have identified as important to advanced problem-solving.

Notes

1. We also ended up interviewing four grade 11 students to confirm that the items could be easily solved by more senior students.
2. Solutions were divided into the three sections, starting point, basic procedure, and final answer, so as to facilitate students' parsing of the information presented. We discovered in pilot tests that when solutions were not divided, students tended to avoid easily the early parts of the solutions.
3. Although 15 students did not complete all sections of both items 1 and 2, we kept the sections they did complete and used the data when we could in analyses. Partly completed items did not interfere with analyses and, more important, we did not want to exclude potentially significant data.
4. The students were assigned to class essentially in a random manner; there was no streaming based on academic ability or prior performance.

References

Cummins, D.D. (1995). Naive theories and causal deduction. *Memory and Cognition, 23,* 646-658.

Downing, C.J., Sternberg, R.J., & Ross, B.H. (1985). Multicausal inference: Evaluation of evidence in causally complex situations. *Journal of Experimental Psychology: General, 114,* 239-263.

Ericsson, K.A., & Simon, H.A. (1993). *Protocol analysis.* Cambridge, MA: MIT Press.

Evans, J.St.B.T. (1989). *Bias in human reasoning: Causes and consequences.* Hillsdale: Erlbaum.

Evans, J.St.B.T., Over, D.E., & Manktelow, K.I. (1993). Reasoning, decision making, and rationality. *Cognition, 49,* 165-187.

Galotti, K.M. (1989). Approaches to studying formal and everyday reasoning. *Psychological Bulletin, 105,* 331-351.

Garnham, A., & Oakhill, J. (1994). *Everyday reasoning. Thinking and reasoning.* Cambridge, MA: Blackwell.

Gigerenzer, G., & Murray, D.J. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Erlbaum.

Hambleton, R.K., & Murphy, E. (1992). A psychometric perspective on authentic measurement. *Applied Measurement in Education, 5,* 1-16.

Hart, D. (1994). *Authentic assessment.* Menlo Park, CA: Addison-Wesley.

Johnson-Laird, P.N., & Byrne, R.M.J. (1991). *Deduction.* Hillsdale, NJ: Erlbaum.

Kulak, A., & Rooney, B. (1999). *The influence of self-report instructions on participants' reported strategies.* Paper presented at a Symposium at the Ninth Annual meeting of the Canadian Society for Brain, Behaviour, and Cognitive Science, Edmonton.

Linn, R.L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher, 23,* 4-14.

Maguire, T.O., Hattie, J., & Haig, B.D. (1993). *Construct validity and achievement assessment.* Paper presented at the Cognition and Assessment Conference, Queen's University, Kingston.

Marascuilo, L.A., & Serlin, R.C. (1988). *Statistical methods for the behavioral sciences.* New York: Freeman.

Newell, A., & Simon, H.A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Norris, S.P. (1983). The inconsistencies at the foundation of construct validation theory. In E.R. House (Ed.), *Philosophy of evaluation* (pp. 53-74). San Francisco, CA: Jossey-Bass.

Norris, S.P. (1985). Competencies as powers. *Philosophy of Education, 40,* 167-178.

Perkins, D. (1986). *Knowledge as design.* Hillsdale, NJ: Erlbaum.

Pollard, P. (1990). Natural selection for the selection task: Limits to social exchange theory. *Cognition, 36,* 195-204.

Robinson, K. (1999). *Development, lies, and videotapes: The validity of children's verbal reports.* Paper presented at the Ninth Annual meeting of the Canadian Society for Brain, Behaviour, and Cognitive Science, Edmonton.

Rogers, W.T., Maguire, T.O., & Leighton, J.P. (1998, July). *Alternative methods for assessing problem solving in junior high school mathematics.* Paper presented at the annual meeting of the Canadian Psychological Association, Edmonton. Abstracted in *Canadian Psychology, 39*:2a, 1998.

Royer, J.M., Cisero, C.A., & Carlo, M.S. (1993). Techniques and procedures for assessing cognitive skills. *Review of Educational Research, 63,* 201-243.

Sugrue, B. (1995). A theory-based framework for assessing domain-specific problem-solving ability. *Educational Measurement: Issues and Practice, 14*(3), 29-36.

Tombari, M., & Borich, G. (1999). *Authentic assessment in the classroom.* Englewood Cliffs, NJ: Prentice-Hall.

Tversky, A., & Kahneman, D. (1982). Availability: A heuristic for judging frequency and probability. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* Cambridge, MA: Cambridge University Press.

Wiley, D.E. (1991). Test validity and invalidity reconsidered. In R.E. Snow & D.E. Wiley (Eds.), *Improving inquiry in social science* (pp. 75-108). Hillsdale, NJ: Erlbaum.