Journal of Educational Psychology
2004, Vol. 96, No. 3, 558–568


0022-0663/04/$12.00 DOI: 10.1037/0022-0663.96.3.558

# Measuring Knowledge to Optimize Cognitive Load Factors During Instruction

Slava Kalyuga and John Sweller
University of New South Wales

The expertise reversal effect occurs when a learning procedure that is effective for novices becomes ineffective for more knowledgeable learners. The authors consider how to match instructional presentations to levels of learner knowledge. Experiments 1–2 were designed to develop a schema-based rapid method of measuring learners' knowledge in a specific area. Experimental data using algebra and geometry materials for students in Grades 9–10 indicated a highly significant correlation (up to .92) between performance on the rapid measure and traditional measures of knowledge, with test times reduced by factors of 4.9 and 2.5, respectively. Experiments 3–4 used this method to monitor learners' cognitive performance to determine which instructional design should be used for given levels of expertise.

The expertise reversal effect (see Kalyuga, Ayres, Chandler, & Sweller, 2003) occurs when an instructional procedure that is relatively effective for novices becomes ineffective for more knowledgeable learners. A consequence of the effect is that an instructor must be able to accurately estimate the knowledge levels of learners to determine an appropriate instructional design for them. Frequently, knowledge levels of learners need to be assessed and monitored continuously during instructional episodes to dynamically determine the design of further instruction. Accordingly, it is critical to have a simple, rapid measure of expertise, especially in computer- and Web-based learning. Current measurement and test procedures may not be adequate for this purpose. The aim of the current work was to devise a rapid test of levels of expertise based on our knowledge of human cognitive architecture and then to use the test as a means of determining instructional procedures.

The expertise reversal effect is an example of an aptitude–treatment interaction (e.g., see Cronbach & Snow, 1977; Lohman, 1986; Snow, 1989) or, more specifically, a disordinal interaction between person characteristics and educational treatment such that if instructional design A is superior to B for novices, B is superior to A for experts. In our research, the expertise reversal effect was derived from longitudinal studies of the effectiveness of different instructional formats and procedures with changing levels of learner expertise and explained using cognitive load theory (see Paas, Renkl, & Sweller, 2003; Sweller, 1999; and Sweller, Van Merrienboer, & Paas, 1998, for reviews), a theory based on the assumption that the processing limitations of working memory might be a major factor influencing the effectiveness of instructional presentations. Working memory capacity is overloaded if more than a few chunks of information are processed simultaneously (see, e.g., Baddeley, 1986; Miller, 1956). To overcome the limitations of working memory, hierarchically organized, domain-specific long-term memory knowledge structures, or schemas, allow people to categorize multiple elements of information as a single higher level element (see Chi, Glaser, & Rees, 1982; Larkin, McDermott, Simon, & Simon, 1980). Because a schema is treated as a single element or chunk, such high-level elements require less working memory capacity for processing than the multiple, lower level elements they contain, making the working memory load more manageable.

Cognitive load imposed by processing instructional material may depend on levels of learner knowledge. For example, textual material initially essential for understanding diagrams may, with increasing levels of knowledge, become redundant. Experts who have acquired considerable high-level schemas in their area of expertise may not require any additional textual explanations. If explanations, nevertheless, are provided, processing this redundant information may increase the load on limited-capacity working memory. (For work on the redundancy effect, see Bobis, Sweller, & Cooper, 1993; Chandler & Sweller, 1991, 1996; Craig, Gholson, & Driscoll, 2002; Mayer, Bove, Bryman, Mars, & Tapangco, 1996; Mayer, Heiser, & Lonn, 2001; Reder & Anderson, 1980, 1982; Sweller & Chandler, 1994.) Kalyuga, Chandler, and Sweller (1998, 2000, 2001), Kalyuga, Chandler, Tuovinen, and Sweller (2001), and Tuovinen and Sweller (1999) found that procedures and techniques designed to reduce working memory overload, such as integrating textual explanations into diagrams to minimize split attention, replacing visual text with auditory narration, or using worked examples to increase levels of instructional guidance, were most efficient for less knowledgeable learners. With the development of knowledge in a domain, such procedures and techniques often became redundant, resulting in negative rather than positive or neutral effects. These redundant sources of information were hypothesized to have imposed an additional cognitive load. Data on subjective rating measures of cognitive load supported these hypotheses. Knowledgeable learners with acquired schemas in a specific area who try to learn relatively new infor-

Slava Kalyuga, Educational Testing Centre, University of New South Wales, Sydney, New South Wales, Australia; John Sweller, School of Education, University of New South Wales, Sydney, New South Wales, Australia.

The research reported in this article was supported by grants from the Australian Research Council to John Sweller and Slava Kalyuga.

Correspondence concerning this article should be addressed to Slava Kalyuga, Educational Testing Centre, University of New South Wales, 12–22 Rothschild Avenue, Roseberry, New South Wales 2018, Australia. E-mail: s.kalyuga@etc.unsw.edu.au

mation in the same area find it more difficult to process diagrams with explanations than diagram-only formats because of the additional, unnecessary information that they must attend to. (See also McNamara, Kintsch, Songer, & Kintsch, 1996, who obtained clear evidence of the expertise reversal effect although they did not interpret their results in a cognitive load framework.)

Why does presenting more experienced learners with well-guided instructions result in a deterioration in performance compared with reduced guidance? Constructing integrated mental representations of a current task is likely to require a considerable share of working memory resources. This activity may be supported either by available schema-based knowledge structures from long-term memory or by external instructional guidance. The relative weight of each component depends on the level of a learner's knowledge in a domain. For novices dealing with novel units of information, instruction may be the only available guidance. For experts dealing with a previously learned familiar domain, appropriate schema-based knowledge can carry out necessary control and regulation functions for the task. Human cognitive architecture dramatically alters the manner in which information is processed as that information increases in familiarity (Sweller, 2003). If more knowledgeable learners are presented instruction intended for schema construction purposes, that redundant instruction may conflict with currently held schemas, resulting in the redundancy and expertise reversal effects. Thus, the optimization of cognitive load in instruction assumes not only the presentation of appropriate information at the appropriate time but also timely removal of inefficient, redundant information as learner levels of knowledge increase.

## Assessing Levels of Expertise

A question of considerable practical interest is how to use the results of studies on the expertise reversal effect and the suggested theoretical explanation of the effect in the design of instructional presentations with an optimized cognitive load. Such instructional designs should be based on user-adapted instructional procedures by matching instructional presentations to levels of learner knowledge in a specific task domain. To achieve this aim, a method of measuring learners' domain-specific knowledge that both is based on knowledge of human cognitive architecture and is quick and reliable is required. Many traditional methods of evaluating learners' knowledge depend on tests involving the solution of a series of problems representing a given task domain. Such methods are usually time consuming and hardly suitable for real-time evaluation of learner progress during instruction required for online adaptation of instructional procedures to changing levels of expertise.

The needs for new approaches to assessment based on cognitive theories have been clearly established, and some promising attempts to link cognitive science and psychometrics have been undertaken (Embertson, 1993; Mislevy, 1996; Pellegrino, Baxter, & Glaser, 1999; Snow & Lohman, 1989; Tatsuoka, 1990). Sophisticated statistical models have been developed and applied to the assessment of multiple cognitive skills involved in the performance of complex tasks (Adams & Wilson, 1996; Embertson, 1991; Martin & VanLehn, 1995; Mislevy, 1994). A schema-based approach to assessment of knowledge (Marshall, 1995; Singley & Bennett, 2002) has not yet resulted in widely usable testing methods.

Most current cognitive assessment theories are aimed primarily at developing new statistical models to apply to the data. However, equally important is the availability of efficient means of gathering evidence of learners' cognitive attributes on which to base statistical inferences (Lohman, 2000). Traditional and newly developed measurement theories may not sufficiently take into account contemporary knowledge of human cognitive architecture and the nature of cognition and instruction. If a major aim of instruction is the construction and storing of schemas in long-term memory and if those schemas alter the characteristics of working memory, then, paradoxically, tests of working memory content may provide a de facto measure of levels of expertise. In effect, such tests should be designed to assess the extent to which working memory limits have been altered by information in long-term memory.

Schemas held in long-term memory, when brought into working memory, effectively form a new memory structure called long-term working memory, which is durable, interference proof, and virtually unlimited in capacity (Ericsson & Kintsch, 1995). When assessing knowledge-based cognitive performance, researchers need to test the content of a student's long-term working memory by using appropriate knowledge-driven tasks. Analyzing the content of long-term working memory during students' cognitive activities can be a powerful way of obtaining valuable evidence about underlying cognitive structures.

Our first approach to developing a rapid cognitive diagnostic test assumed that experts are able to remember more essential elements of the task they encounter than less knowledgeable learners, as demonstrated by the studies of De Groot (1946/1965) and Chase and Simon (1973) on chess expertise. They found that professional grand masters were better able than weekend players to reproduce briefly presented chess positions taken from real games. Schematic knowledge of large numbers of different game configurations held in long-term memory dramatically altered the characteristics of working memory. However, our preliminary studies using coordinate geometry tasks found no significant correlation between learners' ability to reproduce diagrams with a task statement to which they were exposed for several seconds and traditional measures of knowledge based on solving a series of test problems from the same domain. The reproduction task may not have relied sufficiently on cognitive processes associated with problem-solution schemas but rather on images of the diagrams in visual working memory. Isolating the relevant elements of images from the irrelevant elements and then integrating the relevant elements into a solution move may require far larger working memory resources than simply reproducing the diagram.

If knowledge of solution moves associated with a problem state reduces working memory load more than knowledge of the elements of the problem state, then a test of appropriate solution moves may provide a more valid test of expertise than a test that simply emphasizes the elements of the problem state. Such a test could be realized by presenting learners with incomplete intermediate stages of the task solution and asking them to indicate just an immediate next step toward solution, instead of providing all of the solution steps. Because schema-based problem solving requires learners to recognize both problem states and the moves associated with those states, it might be expected that such a measure of expertise might be superior to simply being able to reproduce problem states. More knowledgeable learners presumably should be better able to recognize intermediate problem states and retrieve appropriate solution steps than less knowledgeable learners.

For example, assume the task of solving the algebraic equation $(3x - 5)/2 = 5$. The sequence of main intermediate steps in the solution procedure, corresponding to the subgoal structure of the task, is:

| | |
|---|---|
| Multiply both sides of the equation by 2: | $3x - 5 = 10$. |
| Add 5 to both sides of the equation: | $3x = 15$. |
| Divide both sides by 3: | $x = 15/3 = 5$. |

To be able to solve the original equation, a learner should apply her or his schematic knowledge of solution procedures to all the subtasks. Lack of knowledge of any subtask would interfere with the entire solution procedure. On the other hand, knowing a first move for each subtask leads directly to the next subtask. The testing procedure might be accelerated if, instead of requiring learners to provide complete solutions for all tasks in a test, researchers presented learners with a multilevel series of subtasks for a limited time (e.g., a few seconds for each subtask) with the requirement to indicate the immediate next step toward the complete solution of each subtask. Because the immediate next step for each task level takes a learner to the next task level and because that task level is represented by another task in the series, this rapid testing method may be equivalent to the complete problem-solution alternative.

The following experiments were aimed at developing alternative rapid methods of cognitive diagnosis that emphasized schematic knowledge of solution moves. Experiments 1 and 2 were designed to evaluate the external validity of tests that implemented the above first-step approach. Experiments 3 and 4 tested possible ways of using the methods resulting from Experiments 1 and 2 for building learner-adapted instructional packages that are sourced on monitoring learner performance before and during instruction.

## Experiment 1

Experiment 1 was designed to investigate if a rapid test of learner knowledge in a domain based on a schema theory view of knowledge held in long-term memory could be validated by correlating highly with a multilevel traditional test of knowledge. During the rapid test, students were presented with a set of stand-alone algebraic equations that represented sequential levels of solution of top-level equations—for example, $2(2x - 5)/3 = 5$; $3(x + 1) = 5$; $4x - 5 = 3$; $5x = 3$—and asked to indicate an immediate next step toward solution. More knowledgeable learners presumably should have been better able to recognize problem states and retrieve appropriate moves than less knowledgeable learners. In the traditional test, the same students were required to provide complete solutions of similar multilevel equations. To determine actual time reductions associated with rapid testing in comparison with traditional tests, we used self-paced tasks in both tests.

### *Method*

#### *Participants*

One class of 24 Year 9 (equivalent to Grade 9; 14 girls and 10 boys, approximate age of 15 years) advanced-level mathematics students and one class of 21 Year 10 (equivalent to Grade 10; 10 girls and 11 boys, approximate age of 16 years) intermediate-level mathematics students from a Sydney, Australia, public school located in a lower-middle-class suburb participated in the experiment. Advanced and intermediate mathematics courses differ in their content, levels of difficulty, and allocated study time. The distribution of students between the courses is based on an evaluation of their performance in previous years. By the time of the experiment, all students had completed the sections of the mathematics course necessary for solving the linear equations included in the test.

#### *Materials and Procedure*

The experiment was conducted in a realistic class environment. All participants were tested simultaneously, and all tests were conducted in a single session of about 15 min. The experiment consisted of two tests. The first (traditional) test included a set of 12 equations similar to those described above (three equations at each of four levels). Three equations at the same level were located on each of four pages (e.g., $-4x = 5$, $5x = -4$, and $-4x = -6$ on the first page; $3x - 7 = -3$, $5 - 7x = 4$, and $-4x + 3 = -3$ on the second page; etc.) with the statement "Solve for x" preceding each equation. Students were required to provide complete solutions for all the equations as quickly and accurately as they could and to let the teacher (or experimenter) know as soon as they finished the test. Time taken to complete the test was recorded for each participant. The pages were distributed to students facedown so that all students could start the test simultaneously by turning the pages over. Students' solutions for each equation were assessed as the number of correct steps. Omitted operations (e.g., multiplying out the denominator before canceling out common factors) were counted as separate steps. A total score out of 58 was allocated to each student for the first test.

During the second test, the learners were presented a page containing 12 equations similar to those used in the first test. The equations were arranged in four rows with 3 equations of the same level (with levels defined as before) in each row, starting with the simplest equations in the top row. The distance between rows (about 5 cm) was enough to write answers under each equation. The following statement was placed at the top of the page: "For each of the following equations, indicate the first step (operation) towards the solution." Students were instructed to write down their answers immediately and to let the teacher (or experimenter) know when they finished the test. Time taken to complete the test was recorded for each student. The speed of response was accentuated, so the students' actions were more likely to reflect immediate traces of the content of memory rather than remote results of cognitive processes. The pages were distributed to students facedown so that all students could start the test simultaneously by turning the pages over. Students' answers for each equation were judged as either correct or incorrect, providing a total score out of 12 for the second test. If an answer was not an immediate next step but one of the following steps to the solution, it was counted as a correct answer. It should be noted that the sequence of the test administration was not counterbalanced in this study. Counterbalancing would have eliminated practice effects but caused misinterpretation of correlations between performance scores in the two conditions, which was the main goal of this experiment. Also, counterbalancing would inevitably have increased experimental intrusion into the normally scheduled class learning process.

### *Results and Discussion*

The variables under analysis were traditional test time (time in seconds each learner spent on solving all 12 test tasks; $M = 574.11$, $SD = 164.12$), rapid test time (time in seconds each learner spent on solving all 12 test tasks; $M = 118.22$, $SD = 46.35$), test scores for the traditional test ($M = 33.53$, $SD = 17.61$; 58% correct, actual range of test performance scores from 0 to 58); and test scores for the rapid test ($M = 8.64$, $SD = 3.0$; 72% correct, actual range of test performance scores from 0 to 12). There was a significant difference between test times for the traditional and rapid tests, $t(44) = 20.86$, $p < .01$, Cohen's $f$ effect-size index $= 2.23$, which was expected considering the much smaller number of solution steps that students had to record in the rapid test compared with the traditional test. (For each of the four-level series of equations, the number of steps was reduced from $4 + 3 + 2 + 1 = 10$ to just 4). Students took on average about 10 s per solution step in both conditions, suggesting that practice on the prior, traditional test had minimal influence on the rapid test.

A Pearson product–moment correlation, $r(44) = .92$, $p < .01$, between scores for the traditional and rapid tests was obtained, with a 95% confidence interval extending from .86 to .96, suggesting a very high degree of the concurrent validity for the rapid test. Estimates of Cronbach's coefficient alpha were .78 for the traditional test and .63 for the rapid test. The reliability of a longer test is higher than that of a shorter test. Cronbach's alpha shows how reliably a test measures a single unidimensional latent construct. Considering that our rapid test was designed to measure four separate and distinctive cognitive constructs (associated with schemas for multiplying out the denominator of a fraction, expanding the grouping symbols, adding/subtracting the same number to/from both sides of an equation, and dividing both sides of an equation by the same number), the relatively low (well below the traditionally acceptable 0.8) value of Cronbach's alpha is not surprising. When data have a multidimensional structure, a low Cronbach's alpha is expected. In the traditional test, item scores aggregated different dimensions and thus were less sensitive to the underlying multidimensionality of knowledge structures. A low number of items in both tests might also have contributed to low alphas. However, evidence that the reliability of the rapid test is sufficiently high to allow validity comes from the very high correlation between the two tests.

The results of Experiment 1 indicated a highly significant correlation between learners' performance on the rapid test tasks and traditional measures of learners' knowledge. Furthermore, test time for the rapid method was reduced by a factor of 4.9 in comparison with the time for the traditional test. To further validate this technique, we applied it to a different set of instructional materials in Experiment 2.

## Experiment 2

On the basis of Experiment 1, the rapid test appeared to be a viable candidate to measure levels of expertise for instructional purposes. Experiment 2 was designed to replicate the results of Experiment 1 using coordinate geometry materials.

### *Method*

#### *Participants*

One class of 20 (10 girls and 10 boys) Year 9 advanced-level mathematics students from a Sydney public school (the same class was used in Experiment 1 about 2 months earlier) participated in the experiment. Prior to the experiment, students had been taught sufficient coordinate plane and coordinates of a point geometry to solve the tasks included in the test.

#### *Materials and Procedure*

The experiment was conducted in a realistic class environment. All participants were tested simultaneously, and all tests were conducted in a single session of about 15 min. The experiment consisted of two tests. The first (traditional) test included a set of 12 tasks. Each task included a coordinate plane and two points A and B with given coordinates (see Figure 1). Lines AC and BC were parallel to the *x*- and *y*-axes, respectively. The task was to find the length of AC and BC. The tasks were sequenced according to the level of knowledge they tested (with three tasks for each of four levels). For the highest level tasks, no additional details were provided (see Figure 1, top diagram). For each of the lower levels, progressively more details (indications of coordinates on axes, lines projecting coordinates of the points, etc.) or partial solutions were provided. The bottom diagram in Figure 1 is an example of the lowest level. Three tasks of the same level were located on each of four pages, with the highest level tasks presented on the first page. Students were required to provide complete solutions for all the tasks as quickly as they could and to let the teacher (or experimenter) know as soon as they finished the test. The time taken to complete the test (test time) was recorded for each student. The pages were distributed to students facedown so that that all students could start the test simultaneously by turning the pages over. Students' solutions for each task were assessed as the number of correct steps in the solution, giving a total score out of 30 to be allocated to each student for the first test.

During the second test, the learners were presented four pages containing 12 tasks similar to those used in the first test and sequenced in the same way. Students were required to indicate only the first step toward finding the length of AC and BC. We explained to participants that their answer might include, for example, writing a number or drawing a line on the diagram. Students were instructed to do this test as quickly as they could and to let the teacher (or experimenter) know immediately when they finished the test. Time taken to complete the test was recorded for each student. The pages were distributed to students facedown so that all students could start the test simultaneously by turning the pages over. Students' answers for each task were judged as either correct or incorrect, providing a total score out of 12 for the second test.

In contrast with the algebra equations used in Experiment 1, it was impossible to diagrammatically represent an intermediate state of a geometry task without depicting details of the previous solution steps. The tasks in this experiment could be effectively considered as a sequence of partially worked examples with gradually increasing levels of detail provided to learners. Because the solution details were provided to learners in both the traditional and rapid test tasks, this procedure should not have potentially decreased validity measures. The test sequence was not counterbalanced in this study for the same reasons as in Experiment 1.

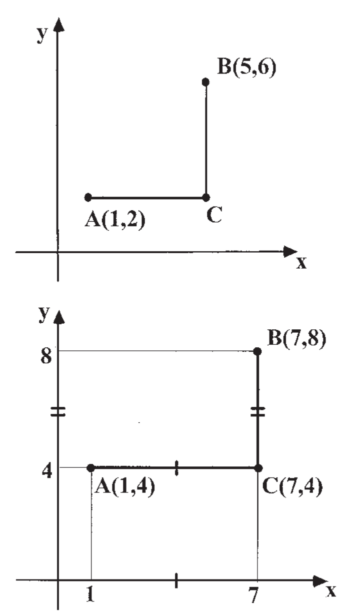


*Figure 1.* Examples of the coordinate geometry tasks used in Experiment 2.

### Results and Discussion

The variables under analysis were traditional test time ($M$ = 306.20 s, $SD$ = 104.97), rapid test time ($M$ = 124.30 s, $SD$ = 24.85), test scores for the traditional test ($M$ = 14.20, $SD$ = 10.62; 47% correct, actual range of test performance scores from 0 to 30), and test scores for the rapid test ($M$ = 4.95, $SD$ = 3.47; 41% correct, actual range of test performance scores from 1 to 12 ). There was a significant difference between test times for the traditional and rapid tests, $t(19) = 8.15$, $p < .01$, Cohen's $f$ effect-size index = 1.33. A Pearson product–moment correlation, $r(19) = .85$, $p < .01$, was obtained between scores for the traditional and rapid tests, with a 95% confidence interval extending from .65 to .94. Estimates of reliability using Cronbach's coefficient alpha were .68 for the traditional test and .57 for the rapid test.

The results of Experiment 2 indicated a high correlation between performance on the rapid test tasks and traditional measures of knowledge requiring complete solutions of corresponding tasks. The test time for the rapid method was reduced by a factor of 2.5 in comparison with the traditional test time. Similar to Experiment 1, students took an average of about 10 s per solution step in both conditions. The same average rate of performance across conditions in both experiments provided us with an indication of the magnitude of response time that could be expected in rapid tests (including retrieval of a solution schema from long-term memory, applying the schema, and recording the result).

Experiments 1 and 2 provided results indicating that basing an achievement test on cognitive theory can generate a test that can be completed very rapidly but that has a high degree of concurrent validity. Experiment 3 was designed to apply that test to predicting which instructional design procedures should be used for students with differing levels of expertise.

## Experiment 3

The next step was to evaluate the ability of the rapid testing procedure to detect experimental effects that had been previously observed using traditional testing methods. One such effect is an interaction between levels of learners' knowledge in a domain and levels of instructional guidance (expertise reversal effect—see Kalyuga et al., 2003). For example, less knowledgeable learners usually benefit from more guided instructional procedures such as worked examples. By contrast, minimal guidance formats (such as solving problems) might be more beneficial for more knowledgeable learners (Kalyuga, Chandler, et al., 2001). The aim of Experiment 3 was to see if we could replicate the expertise reversal effect using the rapid testing technique to assess levels of learners' knowledge in a domain (coordinate geometry) before experimental treatment and to measure levels of learners' performance after their exposure to different instructional formats.

### Method

#### Participants

Two classes (an advanced-level mathematics class and an intermediate-level mathematics class) of 42 Year 9 students from a Sydney Catholic girls' school located in a lower-middle-class suburb participated in the experiment. By the time of the experiment, all students had been taught a basic introduction to the coordinate plane and coordinates of a point geometry necessary for solving the tasks included in the test. The advanced mathematics students had been also taught how to calculate the midpoint of an interval and the distance between two points on a coordinate plane. Those tasks might have provided the learners with some experience in calculating projections of an interval on the coordinate axes, which was the essence of the experimental tasks in Experiment 3. However, the students had not previously encountered tasks formulated using the current format.

#### Materials and Procedure

The experiment was conducted in a realistic class environment. All participants were tested simultaneously, with the experiment conducted in two sessions separated by 1 week.

During the first session (about 5 min long), all participants were presented a rapid test with the purpose of evaluating the initial level of their knowledge in the domain. The test included a set of eight tasks similar to those used in Experiment 2. The following instructions were presented at the top of the first page:

> In each of the figures below, A and B are two points on a coordinate plane. Lines AC and BC are parallel to the coordinate axes. Assume you need to find the lengths of AC and BC.
>
> Some additional details (lines, coordinates) or partial solutions are provided on most figures. For each figure, spend no more than a few seconds to indicate your immediate next step towards solution of the task.
>
> Remember, you do not have to solve the whole task. All you have to do for each figure is to just show the next step towards the solution (for example, it might be just writing a number or drawing a line on the diagram). If you don't know your answer, proceed to the next page.
>
> Do not spend more than a few seconds for each figure and do not go back to pages you have already inspected.

Similar to Experiment 2, the problems were ordered according to the level of knowledge that was required to solve them. In Experiment 3, the levels were more fine-grained than in the previous studies, and instead of four, we used eight levels, with one task corresponding to each level. For the highest level task, no additional details were provided. For each of the lower level tasks, progressively more additional details (indications of coordinates on axes, lines projecting coordinates of the points, etc.) or partial solutions were provided. Each task was presented on a separate page. The test was experimenter paced. After about 10 s on a page, students were instructed to proceed to the next page. Thus, time taken to complete the test was the same for all students (around 90 s).

In Experiment 3, we used a different test-scoring technique from the previous experiments. In the rapid diagnostic tests described above, if an answer was not an immediate next step expected in the fully worked-out, detailed solution but was one of the following steps toward the solution (or even the final step of the solution), it was counted as a correct answer. The same score of 1 was allocated for such an answer as for an answer indicating the immediate next step. However, skipping some intermediate stages of the solution procedure is possible if the learner has corresponding operations automated or is able to perform these operations mentally without writing them down. The ability to skip steps reflects a higher level of knowledge in comparison with the level of knowledge of a learner who can indicate the immediate next step (Blessing & Anderson, 1996; Koedinger & Anderson, 1990; Sweller, Mawer, & Ward, 1983). Knowledge of immediate schematic solution procedures for each separate subtask might not necessarily guarantee the solution of the whole task. Omitting some intermediate steps indicates the student's ability to integrate separate steps in the solution procedure. The rapid diagnostic test scoring method was modified to take into account such differences in learners' knowledge.

In the modified method, if a learner omitted some intermediate stages while trying to find the length of either side of a rectangle, he or she was allocated an additional score for each skipped step. For example, if a participant indicated the final answer for the length of AC on the very first page (skipping three steps), a score of 4 was allocated for this question. An answer consisting of the final step for the length of AC on the second page qualified for a score of 3, and so on. Thus, if a learner was knowledgeable enough to indicate the correct final answers for the length of AC on each of the first four pages and the correct final answers for the length of BC on the following four pages, the allocated (maximum) score was $4 + 3 + 2 + 1 + 4 + 3 + 2 + 1 = 20$.

It might be noted that analyses of students' performance in Experiments 1 and 2 showed that many learners skipped steps during the traditional tests. However, for the analogous rapid tests, the same students often indicated just the immediate next steps without skipping steps. During our pretest instructions to students in Experiments 1 and 2, we did not emphasize that "your immediate next step" was not necessarily the step as determined by a fully detailed, worked-out solution sequence. In Experiment 3, we deliberately explained this point to participants before they commenced the test. Because this explanation was not provided in Experiments 1 and 2, the modified procedure could not be used in those experiments.

On the basis of scores obtained in the rapid test, participants were divided into two groups: more knowledgeable learners (upper median group) and less knowledgeable learners (lower median group). It should be noted that the division did not correspond exactly to the division of advanced and regular mathematics classes. Some higher performers in the regular class received higher scores than lower performers in the advanced class, indicating that using the existing division of classes could not replace the initial rapid diagnostic test for purposes of distributing students between experimental groups. Students in each of these two groups were further randomly allocated to two subgroups according to their performance rank (those with even or uneven performance rank numbers). In the second stage of the experiment, one of these two groups was given worked-examples-based instruction, and the other group was given a problem-solving-based instructional format.

The second stage of the experiment took place 1 week later. Students were assigned to four experimental groups: (a) high knowledge/worked examples (10 students), (b) high knowledge/problem solving (11 students), (c) low knowledge/worked examples (10 students), and (d) low knowledge/problem solving (11 students). Group numbers were not equal because some students who had participated in Stage 1 were absent during Stage 2.

Participants in the problem-solving groups were presented a series of eight problems to solve. The problems were similar to those used in the rapid test in Stage 1 ("A and B are two points on a coordinate plane. Lines AC and BC are parallel to the coordinate axes. Find the lengths of AC and BC"), except that points A and B were located not only in the right upper quarter of the coordinate plane with all the coordinates being positive but could be located in different parts of the coordinate plane with some coordinate numbers being negative.

The worked-examples condition contained a series of four fully worked-out procedures for calculating the lengths of AC and BC. Participants were requested to follow all steps in each example according to a numbered sequence from 1 to 6. Each example was followed by a problem-solving task. The eight tasks used in this condition were identical to the tasks used in the problem-solving condition. The tasks with even numbers were problem-solving tasks identical in both treatments, whereas the tasks with uneven numbers were presented as worked examples in the worked-example condition. Thus, participants in the worked-example condition studied four examples and attempted four problems, whereas participants in the problem-solving condition attempted eight problems. To avoid a split-attention effect, we embedded explanations of procedural steps in worked examples into diagrams as close as possible to the corresponding diagrammatic elements, with arrows used to limit search. All participants in each group were given sufficient time to complete the tasks. It took 12 min to complete this phase of the experiment (those student who finished earlier were encouraged to revise the examples or to check their solutions).

Postinstruction performance levels were again measured using the rapid testing method. The procedure was identical to that used in Stage 1, except that tasks at this stage had points A and B located not only in the right upper quarter of the coordinate plane but in different parts of the coordinate plane, with some coordinate numbers being negative, similar to the instructional condition.

### *Results and Discussion*

A 2 (instructional procedure) × 2 (level of knowledge) analysis of variance (ANOVA) was conducted using the data from Experiment 3. The dependent variable under analysis was postinstruction performance level as determined by the rapid test scores; independent variables were levels of knowledge (high/low) and format of instruction (problems/worked examples).

An analysis of the low-knowledge/high-knowledge main effect produced a significant difference, $F(1, 38) = 25.01$, $MSE = 15.17$, $p < .01$, Cohen's $f$ effect-size index $= 1.58$. As could be expected, high-knowledge learners ($M = 9.10$, $SD = 5.47$; 46% correct, actual range of test performance scores from 0 to 20) performed significantly better than low-knowledge learners ($M = 2.91$, $SD = 2.41$; 15% correct, actual range of test performance scores from 0 to 7). No main effect of experimental formats was found (for the worked-examples group, $M = 6.00$, $SD = 3.69$; 30% correct, actual range of test performance scores from 0 to 15; and for the problem-solving group, $M = 6.00$, $SD = 6.39$; 30% correct, actual range of test performance scores from 0 to 20). It should be noted that according to the expertise measurement scale used in this study, a score of 6 out of 20 does not indicate low knowledge or inability to solve tasks in this domain. Rather, it means a lack of well-learned or automated solution procedures that are typical of the experts in the domain. For example, a person who correctly solves all the test's tasks by consistently using one step at a time without skipping any intermediate operations could only score 8.

Because one of the main purposes of this experiment was to study the change in effectiveness of instructional formats with knowledge, we were primarily interested in the interaction effect between knowledge and instructional procedures. We expected that a difference in relative knowledge in the domain would produce a change in the effectiveness of different methods of instruction. In accordance with this prediction, the interaction data of the 2 × 2 ANOVA were of major interest in this study. There was a significant knowledge–instructional format disordinal interaction for the performance indicator measured by the rapid testing method, $F(1, 38) = 9.04$, $MSE = 15.17$, $p < .01$, Cohen's $f$ effect-size index $= 0.96$, suggesting that the most efficient mode of instruction depends on the level of learners' knowledge.

Following the significant interaction, simple effect tests indicated that for more knowledgeable learners, the problem-solving format produced better results ($M = 10.82$, $SD = 5.79$; 54% correct, actual range of test performance scores from 3 to 20) than worked examples ($M = 7.20$, $SD = 4.64$; 36% correct, actual range of test performance scores from 0 to 15). Although this effect was not statistically significant, $F(1, 19) = 2.46$, $MSE = 27.86$, a Cohen's $f$ index of 0.36 indicated a medium to large effect size. For less knowledgeable learners, the worked-examples group ($M = 4.80$, $SD = 1.99$; 24% correct, actual range of test performance scores from 2 to 8) performed significantly better than the problem-solving group ($M = 1.18$, $SD = 1.08$; 6% correct, actual range of test performance scores from 0 to 3), $F(1, 19) = 27.58$, $MSE = 2.49$, $p < .01$, Cohen's $f$ effect-size index $= 1.20$.

Thus, as the level of knowledge was raised, the performance of the problem-solving group improved more than performance of the worked-examples group. Less knowledgeable learners performed significantly better after studying worked examples. For more knowledgeable learners, there was some indication of problem-solving benefits compared with studying worked examples. A possible floor effect due to the measurement scale used in the study (with an overall mean of 6.0 out of 20) could have reduced levels of statistical significance of the mean differences. Nevertheless, the results demonstrate a strong expertise reversal effect, with levels of expertise determined by the new, rapid test of learners' knowledge.

In Experiment 3, the rapid testing method was used to initially diagnose levels of learners' knowledge in the domain to subdivide the learners into two groups of relative experts and novices. The instructional procedures (worked examples and problem solving) were the same for both novices and experts and were not adapted to the individual levels of expertise. For the next step, we tested the usability of the rapid test as a means of applying real-time individualized adaptation of instructional procedures to current levels of learners' knowledge in a domain.

## Experiment 4

The aim of Experiment 4 was to see if the rapid test could be effectively used in a computer-based training environment for adapting instructional procedures to changing levels of learners' knowledge in a domain. In Experiment 4, the test was used for initial selection of the appropriate levels of instructional materials according to levels of learners' preliminary knowledge in the domain, as well as for monitoring learners' progress during instruction and real-time selection of the most appropriate instructional formats. The learner-adapted instructional procedure was compared with an equivalent procedure without real-time adaptation of instruction to the level of learner knowledge in the task domain (i.e., random relations between learner knowledge and instructional procedures). Elementary algebra was used as the knowledge domain.

### *Method*

### *Participants*

Fourteen Year 10 and 12 Year 11 intermediate-level mathematics students from a Sydney public boys' school located in a lower-middle-class suburb participated in the experiment. At the time of the experiment, students had been taught about algebraic operations and equations necessary for solving tasks included in the test. Year 11 students (for many of whom the materials might have been very simple) were included in the experiment to have participants with different levels of expertise represented in the experimental sample.

### *Materials and Procedure*

All instructions and training for the study were delivered through five Power Macintosh desktop computers. All the computer-based training packages were designed using Authorware Professional (1995). All participants were tested individually in one session (30 to 50 min long). The participants were first presented with general exercises in typing simple algebraic expressions—for example, $2 * x$, $(2x + 1)/3$, and so on. Five tasks were selected to include all operation and bracket signs that learners would encounter in subsequent testing and training sessions. If any element of an expression was typed incorrectly, a learner was invited to correct the error and try again.

The experimental procedure included an initial rapid diagnostic test, an adaptive training session for the experimental group with yoked participants in the control group, and a final rapid diagnostic test. A flowchart of the adaptive procedure for the experimental training session is provided in Figure 2.

*Initial rapid diagnostic test.* After learners completed their exercises in typing algebraic expressions on the computer, they were presented with the initial rapid diagnostic test designed to evaluate their initial level of knowledge in the task domain. The following task statement preceded the test:

> On each of the following three pages, you will see an equation. For each equation, you have to type a single one-line step that you would normally do first when solving the equation on paper.
>
> For example, when asked to solve the equation $2(3x - 1) = 1$, some people would first write $2 * 3x - 2 * 1 = 1$, others could start from $6x - 2 = 1$ or $6x = 3$, and some might even write the final answer ($x = 1/2$) as the first step.
>
> If, when you are given an equation, you do not know how to solve it, click the button "Don't know". You will be allowed no more than one minute to type your answer.

Three equations were presented on the following three pages, each accompanied by the instruction "Type in the first line of your solution and then press *Return*." For the first equation, the system allocated a score of 2 for typing the final answer or 1 for typing an intermediate solution step, e.g., $-3x/3 = 7/3$. A nil score was allocated for a wrong answer not coinciding with any of the available possible correct options or for pressing the *Don't know* button. For the second equation, the system allocated the scores of 4 or 3, respectively, for typing answers at stages similar to those in the first equation, that is, the final step or the step immediately preceding it; a score of 2 for an answer at the level of the first equation (e.g., $4x = -1$ or $4x = 2 - 3$); and a score of 1 for typing an intermediate step on the route to that level (e.g., $4x + 3 - 3 = 2 - 3$). Similarly, for the third equation, the system allocated scores ranging from 6 to 0.

*Training session.* The training session was based on a completion strategy (Van Merrienboer, 1990; Van Merrienboer, Kirschner, & Kester, 2003) realized as a series of faded worked examples (Atkinson, Derry, Renkl, & Wortham, 2000; Renkl & Atkinson, 2003; Renkl, Atkinson, Maier, & Staley, 2002). According to this approach, novices learn most effectively when instructed using fully worked-out examples. As levels of learners' knowledge in the domain increase, parts of the worked examples can be gradually replaced with problem solving (i.e., faded worked examples). The more knowledgeable learners become, the more problem-solving steps and the fewer worked-example steps are included in instruction.

In the learner-adapted format, the allocation of learners to appropriate stages of the faded-worked-examples procedure was based on the outcomes of the initial rapid diagnostic test. Learners who received scores of 0 or 1 for the first equation, no matter what their scores on the other two questions were, started the training phase from the initial, fully worked-out-example level of instruction. Two fully worked-out examples were presented, each followed by a problem-solving exercise. Time spent studying worked examples was user controlled, and time for solving a problem was limited to 3 min. If all attempts within the 3-min limit were unsuccessful, learners were presented with a fully worked-out solution of the problem.

To monitor individual learners' progress, a rapid diagnostic test similar to the first question of the initial diagnostic test was used. If the learner scored 0 on this test, he or she went through a series of six short worked examples for this level of equation. The first two worked examples were fully worked out using an intermediate step, for instance, $4x/4 = 5/4$, whereas the remaining four examples indicated only the final step, for instance, $x = 5/4$. Next, the learner was again presented the rapid diag-

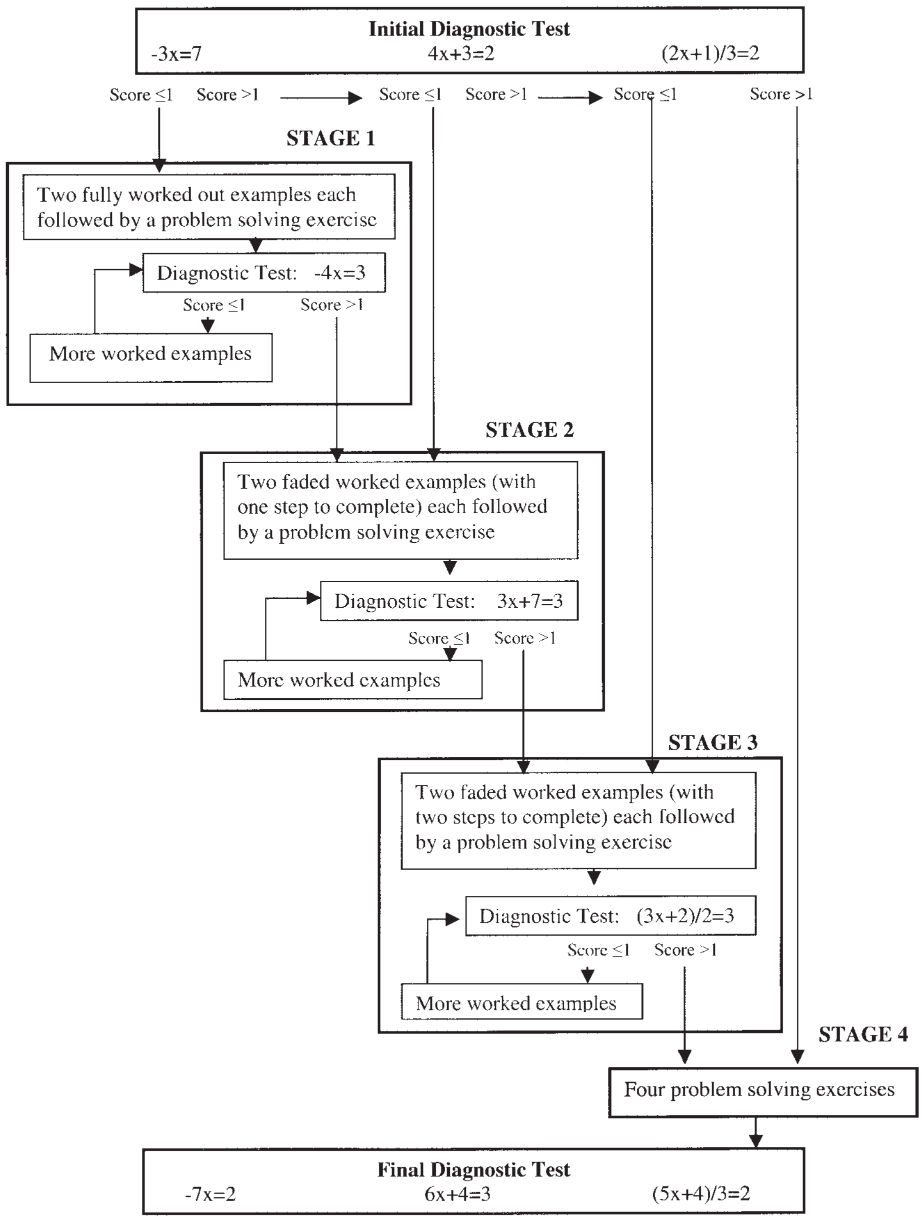


*Figure 2.* Flowchart of the adaptive procedure for the experimental training session.

nostic test for this level. If he or she scored 1, indicating knowledge of the procedure but not enough knowledge to skip the intermediate step, only the last (reduced) four examples were presented. When the learner scored 2 on the rapid test, he or she was allowed to proceed to the next stage of the training session.

The second stage of the training session contained two faded worked examples, each followed by a corresponding problem-solving exercise. In both faded examples, the explanation of the last procedural step (corresponding to the solution of the equation of the type $2x = 5$) was eliminated, and learners were asked to complete the solution themselves and to type in their final answer. If a learner could not solve the remaining equation in 1 min, the correct solution was provided. In problem-solving exercises, similar to the first stage, if learners' attempts within the 3-min limit were unsuccessful, learners were presented with a fully worked-out solution. At the end of the second stage, a rapid diagnostic test similar to the second question of the initial diagnostic test was used. The procedure followed was very similar to the procedure for the first stage.

The third stage of the training session was similar to the second stage except for a lower level of instructional guidance provided to learners (in faded examples, explanations of the two final procedural steps were eliminated) and a higher level of the rapid test at the end of this stage (similar to the third equation of the initial diagnostic test). The fourth and final stage of the training session contained four problem-solving exercises. If learners' attempts to solve each problem within the 3 min limit were unsuccessful, they were presented with fully worked-out solutions.

Thus, in the learner-adapted format, learners who scored 0 or 1 for the first equation of the initial diagnostic test went through all four stages of the training session. How long they stayed at each stage depended on their performance on diagnostic tests during the session. Learners who scored 2 for the first equation of the initial diagnostic test but still scored 0 or 1 for the second equation (no matter what their scores on the third equation were) started the training session from the second stage. Similarly, learners who scored at least 2 for the first and second equations of the initial diagnostic test but scored 0 or 1 for the third equation started the training session from the third stage. Finally, learners who managed to score at least 2 for all three equations of the initial diagnostic test started the training session from the fourth stage, which included only problem-solving exercises.

In contrast, in the non-learner-adapted format, allocation of learners to different stages of the training session was random rather than being based on the results of the initial rapid diagnostic test. To equalize experimental conditions in both groups, each learner in the nonadapted-format group started the training session from exactly the same stage as the previous learner in the learner-adapted-format group. Thus, learners in both groups went through similar stages of the training session. The difference was that in the learner-adapted-format group, the instructional sequence was based on the learner's actual performance on the rapid diagnostic tests. In the nonadapted-format group, the procedure was random in relation to a learner's knowledge. The learner's progress through the training session was not monitored using a rapid diagnostic technique similar to that used in the learner-adapted-format group. Each learner had to study all worked examples and perform all problem exercises that were included in the corresponding stages of the training session, with the same time limits and feedback experienced by the learner's yoked participant.

*Final rapid diagnostic test.* After learners completed the training session, they were presented with the final rapid diagnostic test designed to evaluate their posttraining level of knowledge in the task domain. The test and evaluation procedures were exactly the same as in the initial rapid diagnostic test.

### *Results and Discussion*

The independent variable was the format of the training session (learner adapted or randomly assigned). The dependent variables under analysis were differences between the sum of the three test scores for the final rapid test and the sum of the three test scores for the initial rapid test, providing indicators of learners' knowledge gains due to the training session, and training-session time.

There was a significant difference between groups for knowledge gains, $t(24) = 2.26$, $p < .05$, Cohen's $f$ effect size $= 0.46$. The learner-adapted-format group ($M = 3.23$, $SD = 2.77$) performed significantly better than the randomly assigned–format group ($M = .77$, $SD = 2.77$). There were no significant differences for training-session time ($M = 990.62$, $SD = 353.04$, for the learner-adapted format, and $M = 907.15$, $SD = 426.88$, for the randomly assigned format), $t(24) = .54$, Cohen's $f$ effect size $=$ 0.11. The training-session-time results were expected because of the paired equalization procedure. The significantly higher knowledge gains for the learner-adapted instructional format than the randomly assigned format of training provides strong evidence that the suggested rapid measure of expertise, based on knowledge of human cognitive processes, can be successfully used to enhance learning outcomes by adapting instruction to learners' knowledge levels based on the expertise reversal effect.

Electronic records indicated that students' progress through the learner-adapted instruction depended considerably on their individual performance. For example, 10 (out of 13) participants proceeded through the Stage 1 diagnostic test more than once, 7 participants went through this test more than twice, and 4 participants repeated it more than three times. Ten students proceeded through the Stage 2 diagnostic test more than once, 3 students went through this test more than twice, and 1 student repeated it more than three times. Five participants proceeded through the Stage 3 test more than once, 4 participants went through this test more than twice, and 1 participant repeated it more than three times. Judging by the number of students reattempting tests at different steps, the first operation (dividing both parts of an equation by the same number) was the most difficult one. During training sessions, only 2 students completed all problem exercises and faded worked-examples tasks without errors and repeated attempts; 1 participant reattempted solutions at each of four stages, 3 participants made repeated attempts during at least three stages, 5 students reattempted solutions at two stages, and 2 participants reattempted solutions at one stage. Combined with higher knowledge gains for the learner-adapted group, these records indicate that the suggested tailoring method did individualize instructional procedures as intended.

## General Discussion

If instructional formats and procedures need to change radically with alterations in expertise, a question of considerable practical interest is how to match instructional presentations to levels of learner knowledge. In this article, we have suggested using a rapid method of measuring learner levels of knowledge in a specific area. Students were presented with intermediate stages of a task solution and asked to indicate their next step toward solution for each stage instead of providing a complete solution. Our rationale was that more knowledgeable learners would be able to use their schemas to recognize intermediate problem states and retrieve appropriate solution steps depending on their level of knowledge in the domain. The procedure can be generally described as follows: (a) For a specific task area, establish a sequence of main intermediate steps in the solution procedure corresponding to the subgoal structure of the task; (b) for each step, design representative subtasks, then arrange them in a properly ordered series; and (c) present the series of subtasks to learners for a limited time (e.g., a few seconds for each subtask) with the requirement to quickly indicate the next step toward a complete solution of each task.

Experimental data using algebra and coordinate geometry materials for Year 9 and Year 10 students indicated significant correlations (up to .92) between performance on these tasks and traditional measures of knowledge that required complete solutions of corresponding tasks. Moreover, test times were reduced by factors of 4.9 (for algebra materials) and 2.5 (for coordinate geometry materials) in comparison with traditional test times.

Although correlations between tests were high, measured reliability levels were relatively low. These low values were an artifact of the number and type of test items used. Test reliability increases with increased numbers of items. The essence of effective rapid tests of knowledge is that they have few items but still correlate highly with traditional tests. Nevertheless, low numbers of items result in low reliability indices on normal reliability measures. As well, the possible heterogeneity of content might have contributed to the relatively low internal consistency estimates for the tests in Experiments 1 and 2.

This study has been limited to two narrow domains associated with well-defined tasks and predictable sequences of solution steps (linear algebra equations and simple coordinate geometry). In such areas, the application of the rapid assessment method is straightforward. Establishing the generality of the suggested approach and finding the limits of its usability are important research questions. In more complex domains involving multiple-step problems, students might be able to take many different routes to problem solutions. If all those routes are identifiable, the method still could be used in both paper-based and electronic formats. If the number of routes is too large, a limited number of steps representing different levels of expert-type solutions could be selected for a test. The levels of expertise then could be assessed by, for example, requiring learners to rapidly verify the correctness of each suggested step. To establish the predictive validity of the rapid measures of expertise in other domains, researchers need to test them in other areas of mathematics and science, as well as in less well-structured domains such as text comprehension or second language learning.

Concerning the practical application of the rapid diagnostic procedure in instruction, Experiment 3 confirmed that the suggested testing technique can be used for evaluating levels of learners' knowledge in a realistic learning environment. The technique allowed us to divide learners into appropriate instructional groups and to predict the expertise reversal effect. Experiment 4 indicated that the test could be used as a means of matching different instructional formats to levels of learner knowledge. We used the rapid test to build learner-adapted computer-based instructional procedures based on online monitoring of learner performance before and during instruction. The approach proved to be superior to more traditional approaches. The high values of most effect sizes in Experiments 3 and 4 strongly support the effectiveness of the suggested rapid diagnostic method.

We suggest not only that the rapid diagnostic technique can be used to determine instructional procedures because it is completed rapidly but also that it provides a highly valid cognitive diagnosis because it is designed to directly capture students' schematic knowledge of solution steps. Guiding the sequence and format of instruction should be based on cognitive diagnostic methods. The advantages of doing so can be seen in the results of Experiments 3 and 4.

In our previous studies (Kalyuga, Chandler, & Sweller, 1998, 2000, 2001; Kalyuga, Chandler, et al., 2001), by using subjective ratings of mental load, we obtained evidence for the assumption that redundant information may consume considerable cognitive resources. In future research, we intend to monitor cognitive load during learner-adapted instruction (using, e.g., subjective ratings or a dual-task approach; see Brünken, Plass, & Leutner, 2003; Paas, Tuovinen, Tabbers, & Van Gerven, 2003) to track changes in working memory load during transitions between instructional procedures. By doing so, we may be able to verify the cognitively optimal status of learner-adapted instructional techniques. In this way, the learner-adapted instructional systems could be more efficiently tailored to changing levels of expertise by using combinations of rapid tests of knowledge with measures of cognitive load.

In cognitive diagnostics, the aim of the test is to assess learners' individual cognitive structures rather than their overall level of performance in a domain, as occurs using traditional tests. This procedure is an example of the use of multidimensional assessment tasks, and appropriate multidimensional measurement models need to be applied to the data to make valid statistical inferences. Exploring different multidimensional measurement approaches (e.g., based on multidimensional item-response theories or Bayesian inference networks) in conjunction with rapid testing techniques should be another important direction of future research.

Knowledge of human cognitive architecture and processes has advanced substantially over the past 3 or 4 decades. Although that knowledge has been used by, for example, cognitive load theory to devise novel instructional procedures, some of those procedures are critically dependent on researchers' being ably to quickly and accurately measure learners' levels of expertise. However, no appropriate, cognitively oriented expertise assessment procedures are available to be used in conjunction with the new instructional designs that are rapidly appearing. This article is intended as a first step in remedying this deficiency.

## References

Adams, R. A., & Wilson, M. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard, Jr., & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 143–166). Norwood, NJ: Ablex Publishing.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. W. (2000). Learning from examples: Instructional principles from the worked example research. *Review of Educational Research, 70,* 181–214.

Authorware Professional 3.0 [Computer software]. (1995). San Francisco: Macromedia.

Baddeley, A. D. (1986). *Working memory*. New York: Oxford University Press.

Blessing, S. B., & Anderson, J. R. (1996). How people learn to skip steps. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 576–598.

Bobis, J., Sweller, J., & Cooper, M. (1993). Cognitive load effects in a primary school geometry task. *Learning and Instruction, 3,* 1–21.

Brünken, R., Plass, J., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist, 38,* 53–61.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8,* 293–332.

Chandler, P., & Sweller, J. (1996). Cognitive load while learning to use a computer program. *Applied Cognitive Psychology, 10,* 1–20.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4,* 55–81.

Chi, M., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (pp. 7–75). Hillsdale, NJ: Erlbaum.

Craig, S., Gholson, B., & Driscoll, D. (2002). Animated pedagogical agents in multimedia educational environments: Effects of agent properties, picture features, and redundancy. *Journal of Educational Psychology, 94,* 428–434.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interaction.* New York: Irvington Publishers.

De Groot, A. D. (1965). *Thought and choice in chess.* The Hague, the Netherlands: Mouton. (Original work published 1946)

Embertson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56,* 495–516.

Embertson, S. (1993). Psychometric models for learning and cognitive processes. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 125–150). Mahwah, NJ: Erlbaum.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102,* 211–245.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38,* 23–31.

Kalyuga, S., Chandler, P., & Sweller, J. (1998). Levels of expertise and instructional design. *Human Factors, 40,* 1–17.

Kalyuga, S., Chandler, P., & Sweller, J. (2000). Incorporating learner experience into the design of multimedia instruction. *Journal of Educational Psychology, 92,* 126–136

Kalyuga, S., Chandler, P., & Sweller, J. (2001). Learner experience and efficiency of instructional guidance. *Educational Psychology, 21,* 5–23.

Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of Educational Psychology, 93,* 579–588.

Koedinger, K. R., & Anderson, J. R. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. *Cognitive Science, 14,* 511–550.

Larkin, J., McDermott, J., Simon, D., & Simon, H. (1980). Models of competence in solving physics problems. *Cognitive Science, 4,* 317–348.

Lohman, D. F. (1986). Predicting mathemathanic effects in the teaching of higher-order thinking skills. *Educational Psychologist, 21,* 191–208.

Lohman, D. F. (2000). Complex information processing and intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence* (2nd ed., pp. 285–340). Cambridge, MA: Cambridge University Press.

Marshall, S. (1995). *Schemas in problem solving*. New York: Cambridge University Press.

Martin, J., & VanLehn, K. (1995). A Bayesian approach to cognitive assessment. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 141–165). Hillsdale, NJ: Erlbaum.

Mayer, R., Bove, W., Bryman, A., Mars, R., & Tapangco, L. (1996). When less is more: Meaningful learning from visual and verbal summaries of science textbook lessons. *Journal of Educational Psychology, 88,* 64–73.

Mayer, R., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology, 93,* 187–198.

McNamara, D., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14,* 1–43.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59,* 439–483.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33,* 379–416.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38,* 1–4.

Paas, F., Tuovinen, J., Tabbers, H., & Van Gerven, P. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38,* 63–71.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the “two disciplines” problem: Linking theories of cognition and learning with assessment and instructional practice. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (Vol. 24, pp. 307–353). Washington, DC: American Educational Research Association.

Reder, L., & Anderson, J. R. (1980). A comparison of texts and their summaries: Memorial consequences. *Journal of Verbal Learning and Verbal Behavior, 19,* 121–134.

Reder, L., & Anderson, J. R. (1982). Effects of spacing and embellishment on memory for main points of a text. *Memory & Cognition, 10,* 97–102.

Renkl, A., & Atkinson, R. K. (2003). Structuring the transition from example study to problem solving in cognitive skills acquisition: A cognitive load perspective. *Educational Psychologist, 38,* 15–22.

Renkl, A., Atkinson, R. K., Maier, U. H., & Staley, R. (2002). From example study to problem solving: Smooth transitions help learning. *Journal of Experimental Education, 70,* 293–315.

Singley, M. K., & Bennett, R. E. (2002). Item generation and beyond: Applications of schema theory to mathematics assessment. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 361–384). Mahwah, NJ: Erlbaum.

Snow, R. E. (1989). Aptitude-treatment interaction as a framework for research on individual differences in learning. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences: Advances in theory and research* (pp. 13–59). New York: Freeman.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. Linn (Ed.), *Educational measurement* (pp. 263–331). New York: Macmillan.

Sweller, J. (1999). *Instructional design.* Melbourne, Australia: Australian Council for Educational Research.

Sweller, J. (2003). Evolution of human cognitive architecture. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 43, pp. 215–266). San Diego, CA: Academic Press.

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction, 12,* 185–233.

Sweller, J., Mawer, R., & Ward, M. (1983). Development of expertise in mathematical problem solving. *Journal of Experimental Psychology: General, 12,* 639–661.

Sweller, J., Van Merrienboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10,* 251–296.

Tatsuoka, K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–487). Hillsdale, NJ: Erlbaum.

Tuovinen, J., & Sweller, J. (1999). A comparison of cognitive load associated with discovery learning and worked examples. *Journal of Educational Psychology, 91,* 334–341.

Van Merrienboer, J. J. G. (1990). Strategies for programming instruction in high school: Program completion vs. program generation. *Journal of Educational Computing Research, 6,* 265–287.

Van Merrienboer, J. J. G., Kirschner, P. A., & Kester, L. (2003). Taking the load off a learner’s mind: Instructional design principles for complex learning. *Educational Psychologist, 38,* 5–13.

Received October 13, 2003
Revision received March 1, 2004
Accepted March 31, 2004 ■