When Less Is More in Cognitive Diagnosis: A Rapid Online Method for Diagnosing Learner Task-Specific Expertise

Slava Kalyuga
University of New South Wales

Rapid cognitive diagnosis allows measuring current levels of learner domain-specific knowledge in online learning environments. Such measures are required for individualizing instructional support in real time, as students progress through a learning session. This article describes 2 experiments designed to validate a rapid online diagnostic method that was inspired by experimental procedures applied in classical cognitive studies of chess expertise. With the described rapid verification method, learners are required to rapidly verify suggested steps at various stages of a problem solution procedure. In this study involving 33 university students, a high degree of correlation was found between rapid testing scores and results of in-depth cognitive diagnosis based on observations of problem-solving steps using video recordings and concurrent verbal reports in the domains of kinematics (vector addition motion problems) and mathematics (transforming graphs of linear and quadratic functions). The article discusses possible applications of the suggested method in adaptive learning environments.

Keywords: online cognitive diagnosis, rapid diagnostic method, expertise, learner-tailored instruction

Task-specific expertise is the ability of a person to perform fluently in a specific class of tasks. It has a more narrow definition than professional expertise (e.g., Ericsson & Charness, 1994), and it is an essential part of academic domain-specific expertise within the model of domain learning (Alexander, 2004). For example, a secondary school student could reach a top level of task-specific expertise in solving simple linear algebra equations, although he or she still would be far from becoming not only an expert mathematician but also an expert in school-level mathematics. Although both professional expertise and academic domain learning expertise include additional essential attributes (e.g., a systemic vision of the field, well-developed metacognitive skills, strategic processes, attitudes and interests), a common major feature of all types of expertise is the availability of a well-organized domain-specific knowledge base (Bransford, Brown, & Cocking, 1999). In the case of task-specific expertise, it is knowledge structures (including strategies and procedures) used in solving a specific class of tasks. Because of the full or partial automation of such structures, experts are able to rapidly perform advanced solution stages by integrating procedures and skipping some (or all) intermediate steps (Blessing & Anderson, 1996; Koedinger & Anderson, 1990; Sweller, Mawer, & Ward, 1983).

Task-specific expertise is a necessary prerequisite of both high-level professional expertise and academic domain learning expertise. Experts of both these types are also highly proficient in solving key classes of tasks within their areas of expertise. Therefore, in education and training, developing expertise in main classes of tasks is an important condition of mastering specific subject domains. Within a cognitive load framework (see Van Merriënboer & Sweller, 2005, for a recent overview), the importance of developing task-specific expertise is explained by the need to free cognitive resources that are required for learning higher level tasks and developing flexible and transferable skills.

A number of recent studies have demonstrated significant interactions between levels of learner task-specific expertise and optimal instructional formats (expertise reversal effect; for overviews, see Kalyuga, 2005; 2006b; Kalyuga, Ayres, Chandler, & Sweller, 2003). Many instructional procedures and formats that are effective for novice learners may become ineffective, or even detrimental, for expert learners and vice versa. The major instructional implication of the effect is the need to adapt the design of learning environments and levels of instructional guidance to changing levels of learner expertise in a corresponding class of tasks. To be able to do this dynamically, in real time (e.g., during a single tutoring session), appropriate rapid methods of cognitive diagnosis are required that are capable of accurately detecting different levels of learner task-specific expertise. Such rapid methods are especially important for adaptive online learning environments and Web-based courses that cannot use traditional paper-based diagnostic instruments (or their online equivalents) in real time. E-learning courses and tutorials are becoming widespread; however, their adaptive capabilities are usually limited to tailoring instructional content to relatively superficial learner attributes (e.g., preferences, interests, choices, history of previous online behavior) and are not based on fundamental cognitive characteristics such as learner knowledge base.

Traditional methods that are best suitable for diagnosing individual levels of expertise in a domain are based on interviews and think-aloud protocols as learners perform domain-specific tasks. Such verbal reports may provide near-direct evidence about real cognitive processes, ways of reasoning, and underlying learner knowledge structures. However, these mostly laboratory tools are not suitable for the real-time online evaluation of learner progress during instruction. On the other hand, traditional educational tests either have limited diagnostic capabilities or are too time consum-

I thank Lauren Willis for her assistance in collecting experimental data. Correspondence concerning this article should be addressed to Slava Kalyuga, School of Education, University of New South Wales, Sydney, New South Wales, 2052, Australia. E-mail: S.kalyuga@unsw.edu.au
A Rapid Testing Approach in Cognitive Diagnosis

WM is a setting for people’s conscious cognitive processes. It is severely limited in duration and capacity when dealing with unfamiliar information (Baddeley, 1997; Miller, 1956; Peterson & Peterson, 1959). However, in well-known domains, the available knowledge base in LTM allows people to encapsulate many elements of information into larger chunks that are treated as single units in WM. Available LTM knowledge structures essentially define the characteristics of WM by effectively extending its capacity and reducing WM load. Therefore, WM capacity is not a stable characteristic of a learner but always depends on the learner knowledge base and directly reflects the available domain-specific knowledge structures. Accordingly, a diagnostic method based on immediate evaluation of units of WM content when an individual approaches a task could be used for measuring levels of expertise in the corresponding task domain. The method would actually evaluate the extent to which WM limits have been altered by domain-specific knowledge structures held in LTM.

Because of the association with stable LTM structures, the content of experts’ knowledge units in WM is sufficiently durable and resistant to temporary interferences (Ericsson & Kintsch, 1995) to allow a working diagnostic procedure. If a person has a well-learned strategy for solving a specific class of tasks, he or she will immediately apply it when encountered with a task from this class. Someone without such knowledge will start searching for a solution randomly. It is practically possible to determine the content of knowledge (if any) used in a specific task situation, for example, by analyzing the concurrent (think-aloud) verbal reports. This method, however, is time consuming and difficult to use in online learning environments.

Some experimental procedures used in classical cognitive studies of chess expertise by De Groot (1965) and Chase and Simon (1973) suggested another idea. In those studies, professional grand masters performed considerably better than weekend players in reproducing briefly presented (for few seconds) chess positions taken from real games. There were no significant differences, however, for random configurations of chess figures. A schematic knowledge base of a huge number of different real game configurations held in grand masters’ LTM allowed them to reproduce chess positions by familiar chunks rather than by individual chess figures without overloading WM.

Chess grand masters know the best move for each of the remembered typical real game configurations. Therefore, a rapid test of chess expertise could be based on presenting typical game configurations for brief periods of time and asking players to rapidly indicate their first move for each situation (essentially these are tasks that grand masters face when playing simultaneously on many chess boards). Applying this idea to the rapid diagnostic assessment of task-specific expertise in school mathematics resulted in the first-step diagnostic method (Kalyuga & Sweller, 2004). Learners were presented for a limited time with a series of tasks with gradually changing levels of complexity and were asked to rapidly indicate their first step toward solution of each task. The first step would involve different cognitive processes for individuals with different levels of task-specific expertise. For example, experts would bring their well-automated higher level procedures that allow them to rapidly generate advanced stages of the solution or even final answers to relatively more difficult tasks and skip many intermediate operations. On the other hand, complete novices may only be able to start the first attempt in their random search (e.g., trial-and-error) approach to even simple tasks. Therefore, different first steps would indicate not only the availability but also different levels of acquisition of corresponding knowledge structures in the learner’s LTM.

Validation studies of the first-step method in algebra, coordinate geometry, and arithmetic word problem solving indicated significant correlations (.85, .92, and .72, respectively) between performance on these tasks and traditional measures of knowledge that required complete solutions of corresponding tasks (Kalyuga, 2006c; Kalyuga & Sweller, 2004). Test times were reduced by factors of up to 4.9 in comparison with traditional test times. The first-step method was sensitive to underlying knowledge structures and was sufficiently rapid. The method was used for optimizing levels of instructional guidance in adaptive computer-based tutoring in solving linear algebra equations (Kalyuga & Sweller, 2004, 2005). At the beginning of a session, each learner was allocated to an appropriate level of guidance according to the outcome of the initial rapid diagnostic test. Depending on the outcomes of the rapid diagnostic probes during instruction, the learner was allowed to proceed to the next stage of the session or was required to repeat the same stage and then take the rapid test again. At each subsequent stage of the tutoring session, a lower level of instructional guidance was provided to learners, and a higher level of the rapid diagnostic tasks was used at the end of the stage.

Rapid Verification Method

Another possible approach to a rapid evaluation of chess expertise may be based on presenting a real game configuration for a brief period of time, followed by displays of several possible (both suitable and unsuitable) moves for this configuration, one display at a time. A player should rapidly verify the suitability of each of these moves. A similar approach to the rapid diagnostic assessment of task-specific expertise could also be applied in an online educational context. Learners could be presented with a series of possible (correct and incorrect) steps corresponding to various stages of the task solution procedure and asked to rapidly verify the suggested steps. Specific forms of responses could be clicking on displayed on-screen buttons or pressing specified keys on the computer keyboard (e.g., correct, incorrect, or don’t know). An
The ability to rapidly verify advanced stages of a solution procedure is based on the effectively expanded WM capacity and reduced cognitive load due to available task-relevant knowledge structures (e.g., procedures, rules). The levels of acquisition of this knowledge could be measured by an appropriately defined scoring method that reflects different levels of performance in verifying suggested solution steps. The required rapidness of learners’ responses is not only a means of reducing testing time. More importantly, it is essential for capturing knowledge structures that learners actually use while approaching a task and before any lengthy chains of reasoning or searching could be applied, thus diagnosing the level of knowledge-based expertise.

The rapid diagnostic approach could be considered as a form of dynamic assessment that measures students’ ability to move on with a task solution given a certain level of additional guidance (Bransford & Schwartz, 1999). In the case of rapid verification tasks, the additional guidance is provided by the depiction of a gradually changing number of previously completed steps. When presenting learners with various stages of a solution procedure for rapid verification, researchers use the varying levels of scaffolding to determine the level of learner proficiency in handling increasingly difficult situations.

The rapid verification method was actually used for optimizing levels of instructional guidance and individualizing instructional procedures in an adaptive computer-based tutorial in kinematics (Kalyuga, 2006a). As levels of learner expertise increased according to online rapid verification tests, the levels of provided guidance were reduced. On the other hand, for learners with insufficient levels of knowledge, more guidance was provided in the form of worked-out solution steps. However, in contrast to the first-step method, no validation studies were conducted for this method. Test scores are valid if they are obtained from a measure designed, or method implemented, to evaluate an attribute, construct, or variable that has been defined clearly on the basis of a theory and substantive reviews of literature. Using correlations of the test scores with other measures of the attribute as evidence that the test assesses the intended attribute is referred to as establishing the concurrent validity of the test. Thus, investigating whether the rapid assessment results would correlate highly with more traditional measures of task-specific expertise is needed to support the claim that the method actually measures the levels of learner expertise.

For example, there could be concerns that replacing the first-step rapid diagnostic method with a rapid verification technique would effectively turn a test of organized knowledge structures into a recognition test measuring knowledge of shallow task characteristics. The rapid verification method actually uses a recognition test format for verifying suggested solution steps. However, learners need to recognize intermediate advanced steps in a solution procedure, and these steps have to be rapidly constructed first by retrieving and using available knowledge in LTM. These processes involve more complex cognitive activities and higher levels of knowledge than those required by traditional recognition tests. Therefore, it is assumed that this method is diagnostically more powerful than simple recognition tests and could be used for measuring levels of learner task-specific expertise. However, this assumption needs to be supported by studies of concurrent validity of the diagnostic procedure.

Thus, the experimental studies reported in this article were designed to answer the following research questions:

1. Is the suggested online verification procedure a valid diagnostic method capable of detecting different levels of acquisition of domain-specific knowledge structures? The hypothesis is that the rapid verification test results will correlate significantly with alternative traditional measures of task-specific expertise, thus demonstrating a high degree of concurrent validity.

2. Is the suggested online verification procedure a rapid method that could be completed fast enough for real-time applications? The hypothesis is that the rapid verification test will allow a substantial reduction of diagnostic assessment time in comparison with alternative traditional measures of task-specific expertise.

3. Will the suggested rapid verification procedure generalize to different task domains? The hypothesis is that the rapid tests will demonstrate high degrees of validity indicators in different task areas.

In addition to validating a new diagnostic method that implements the rapid diagnostic approach, the novelty of this study is also in using task domains (vector addition problems and graph transforming tasks) that essentially rely on graphical representations of task situations. Most previous studies of the rapid first-step diagnostic technique used numerical-only tasks. Another essential improvement is observing participants’ actual problem-solving steps using video recordings and concurrent verbal reports as better criterion measures of task-specific expertise than students’ records of problem-solving steps used in previous studies. Thus, in the traditional paper-based test, participants were required to provide complete solutions of tasks similar to those used in the rapid verification test. The participants’ on-paper actions and think-aloud verbalizations were recorded and analyzed. In order to determine actual time reductions associated with rapid testing in comparison with the traditional test, I used self-paced tasks in both tests.

In both test conditions, more knowledgeable learners were expected to perform their tasks with lower levels of cognitive load than relative novices because of effectively increased WM capacity due to the available knowledge base. Therefore, I also included the evaluation of cognitive load in the procedure to provide another indicator of levels of learner expertise in addition to the task performance scores. Previous cognitive load research studies have indicated that a simple subjective rating scale of task difficulty could be an effective means of measuring cognitive load imposed by instructional materials (e.g., see Paas, Tuovinen, Tabbers, & van Gerven, 2003, for an overview).

If the tests actually measure levels of learner task-specific expertise, difficulty ratings are expected to show significant negative correlations with test scores: Because of reduced WM load, tasks in both tests should be relatively easier for more proficient learners than for less experienced participants. In this study, the measures of cognitive
load may also provide an indicator of participants’ actual engagement in knowledge-based higher level cognitive activities during rapid assessment tasks. If the rapid tests were based only on simple recognition of surface task characteristics, the task difficulty would not differ for learners with different levels of expertise. Consequently, there would be no correlations between difficulty ratings and test scores for rapid verification tasks.

Experiment 1

Experiment 1 was designed to investigate whether rapid verification scores in a task domain in kinematics could be validated by correlating highly with the results of observations of participants’ paper-based problem solving steps and their concurrent verbal reports. The experimental tasks represented a class of problems in kinematics called vector addition motion problems. A typical task in this domain requires adding two vectors that are positioned at a certain angle to each other, for example, “A sea wave is traveling at 8 m/s towards the beach. A swimmer moves at 3 m/s in a direction perpendicular to the direction of the wave. What is the velocity of the swimmer relative to the ground?” During the rapid verification test, students were presented with a set of possible (correct and incorrect) intermediate solution steps and were asked to rapidly verify the correctness of these steps. More knowledgeable learners presumably should be better able to rapidly recognize more advanced intermediate solution moves than less knowledgeable learners.

For example, a person who knows that a vector approach should be applied but who has not practiced graphical addition of vectors may be able to verify correctly only a diagram with two perpendicular vectors as a valid step toward the solution. An individual who has more experience with vectors may rapidly verify perpendicular vectors with numerical values assigned to the length of each vector. Another person who is familiar with the vector addition procedure may also verify immediately a diagram representing the graphical addition of these vectors. Someone with more experience in adding vectors might be able to rapidly verify a numerical expression for the Pythagorean theorem that is used in solving this class of tasks. A learner with substantial experience in solving such tasks may even be able to verify a numeric expression representing the final answer without a diagram present.

Method

Participants

Thirty-three university students (18 women and 15 men, age 18–25) participated in this experiment. They represented different years of study: 17 undergraduate, 3 graduate, and 13 postgraduate students. Participants also represented variety of subject areas: 12 students were involved in technical areas (including mathematics, mechanical engineering, computer engineering, biotechnology), and 21 were from nontechnical areas (education, psychology, medicine, law, management, international studies). In order to have a wide range of students’ familiarity with the domain and different levels of their task-specific expertise, I deliberately recruited the participants from various areas of academic specialization. With different degrees of involvement of mathematics. Brief pretest interviews indicated that participants represented different levels of expertise in the area of kinematics ranging from novices (some still remembered that they had studied related material in their high school science courses) to experienced individuals studying university engineering and mathematics courses. The scores obtained from the rapid verification tasks and traditional measures of expertise in the experiment confirmed the initial pretest rough evaluations. Students with more expected experience in the domain performed better on the experimental tasks. For example, according to the values of rapid test scores (out of the maximum possible total score of 75), 2 participants representing nontechnical fields were in the range from 0 to 25 (the lower third), another 19 nontechnical participants were in the range from 26 to 50 (the medium third), and 11 participants representing technical areas were in the range from 51 to 75 (the upper third). The students had not been exposed to the specific materials used in the study prior to the experiment. They were paid AU$20 for their participation in the experiment.

Materials and Procedure

Each participant was tested individually in a laboratory environment. Computer-based test items (designed using Authorware Professional; Macromedia, 2003) were delivered through a laptop personal computer. The experimental procedure included a rapid computer-based diagnostic test and a paper-based test with recording of students’ on-paper actions and think-aloud verbal reports. The sequence of the test administration was counterbalanced: Approximately one-half of students performed the rapid test first, and the rest performed the paper-based test first.

Rapid diagnostic test. The test included five tasks corresponding to the following values of angles between vectors: 0° (the same direction of movements), 180° (opposite directions of movements), 90° (perpendicular vectors), 120°, and 60°. In addition, when 60° or 120° angles were used, only equal velocity values for both vectors were allowed. Without these restrictions, the procedures for calculating the length of the resulting vectors required more advanced knowledge of trigonometry that was not a part of the assessment objective in this study. Each textual task statement was followed by five suggested solution steps (correct or incorrect) for rapid verification.

The first verification subtask for each task provided vector graphs indicating only directions of movements. The second verification subtask provided vector graphs with velocity values indicated next to them. For example, for the previously mentioned task that described a situation with perpendicular directions of movements, vector graphs indicating perpendicular directions of movements with corresponding velocity values were provided. The third verification subtask, in addition to the vectors and their values, graphically represented the vector addition operation. For example, for the fourth task (“A boat is traveling at 5 m/s. A passenger runs across the deck at 5 m/s in a direction of 120° relative to the direction of motion of the boat. What is the velocity of the passenger relative to the water?”), the third verification subtask is presented in Figure 1 (incorrect step). The fourth verification subtask provided all necessary graphical information and indicated a numerical expression for calculating the length of the resulting vector. For example, for the above (120° angle) task, a simple expression V = 5 m/s was placed next to the diagram (60° angles and equal sides in two equilateral triangles were also indicated on the diagram). Finally, the fifth verification subtask
included final numerical answer (an integer or surd) with no graphics provided.

Students were instructed that each task in the test would be displayed for a limited time and that following each task, several possible (correct and incorrect) solution steps would be presented one at a time. The task exposure time (15 s) was established in preexperimental trials as sufficient for reading and comprehending task statements. Students were told that most of the suggested steps were possible intermediate stages on a way to the solution, but some suggested steps could also indicate final answers. For each suggested step, students had to immediately verify whether this could be a correct step leading to the solution (or providing the final answer). Each solution verification window included a diagrammatic and/or numerical representation of a possible solution step and the buttons right, wrong, and don’t know for students to click on.

Before the rapid test, the participants were coached in responding sufficiently rapidly using exercises with simple tasks from a different area (common arithmetic calculations). During those exercises, the students got a sense of what was considered a rapid response. If a student did not respond within a set short time interval of few seconds, she or he was asked to respond faster next time and was recommended to try another exercise task. This brief procedural pretraining session was chosen instead of mechanically limiting the allowed verification response time to several seconds. The system-controlled response time (by automatically switching to the next verification window or task after a fixed short time interval) could forcefully interrupt genuine verification responses, thus invalidating results.

Each rapid diagnostic task was followed by a subjective rating of task difficulty, with a 9-point scale ranging from 1 (extremely easy) to 9 (extremely difficult) presented on the screen for students to click on. Students’ response times, performance scores, difficulty ratings, and test times were automatically recorded by the software. In this experiment, the scores allocated for correct responses depended on the level of advancement of intermediate solution stages represented in different verification subtasks. For example, the first subtask required learners to verify the application of only one step (a graphical representation of vectors), and a score of 1 was allocated for a correct response. On the other hand, the fifth subtask required learners to verify the result of the application of five sequential procedural steps, and a score of 5 was allocated for a correct response. Null scores were allocated for incorrect responses and “don’t know” entries. Prior to the test, participants were instructed to enter “don’t know” instead of guessing their responses any time they were in doubt. Because this was not a high-stake test, the students were expected to follow these instructions. Therefore, both incorrect responses and “don’t know” entries were assumed to reflect a lack of specific knowledge, and null scores were allocated for both types of responses. The maximum possible total score for all five tasks in this test could be $5 \times (5 + 4 + 3 + 2 + 1) = 75$.

Paper-based test. Five textual task statements presented to learners were similar to those used in the rapid test (wording and numerical values in the task statements were changed). Each task statement was printed at the top of a separate page. Students were instructed to provide a complete solution for each task below the task statement as quickly as they could and think aloud while solving the task. Before commencing the procedure, participants were briefly coached in how to deliver a think-aloud protocol. The moderator (a research assistant) instructed them to think out loud at all times (e.g., “It really helps me to understand what you are thinking while solving the problem. If you get quiet, I will ask you to keep talking”). The moderator prompted participants to continue thinking aloud (“Please keep talking”) every time they were silent for more than several seconds. Each diagnostic task was followed by a paper version of the subjective rating scale of task difficulty. The scale was identical to that used in the computer-based test; however, instead of clicking on their responses, students were asked to tick or circle them.

Time taken to complete the test was recorded for each participant. Students’ problem-solving steps on paper as well as their verbalizations were recorded using a digital video camera. The camera was mounted on a tripod and focused on paper pages. Only the pages and students’ hands were visually recorded. The research assistant collected the audiovisual performance records for each participant, and I subsequently analyzed them in order to evaluate levels of student task-specific expertise. The most important performance indicators that I looked for during this analysis were the following: search processes (if any) performed prior to executing specific solution steps, time taken by the participant before starting to carry out specific solution steps, and consistency and continuity of the solution process. For example, if the records demonstrated that a participant spent time on searching for possible moves at the beginning and/or subsequent stages of the solution process (rather than immediately applying available knowledge of relevant solution steps), her or his perceived level of expertise in this task domain was accordingly lowered.

A student’s performance in each task was quantitatively assessed as the number of correct solution steps that the student had completed continuously within a short period of time (usually 10–20 s) of starting the solution. This score was determined on the basis of the analysis of both visual and audio recordings of the student’s actions. The steps that were preceded by long chains of reasoning or searching and required more time did not count (even if they were eventually
completed correctly), because they were not based on immediately available knowledge of solution procedures in LTM. Thus, because five major steps were required to complete each of the five tasks, scores from 0 to 5 were allocated for a task. A total score out of 25 was assigned to each participant. The validity of this scoring method was supported by evident associations between the scores and expected levels of expertise based on participants’ academic specializations: Students studying mathematics and technical subjects generally scored higher than those in nontechnical areas. For example, 11 out of 12 participants representing technical areas and only 5 out of 21 representing nontechnical areas were in the upper half (with scores above the mean).

Results and Discussion

Data for 1 student were lost because of a software problem. The variables under analysis were as follows: paper-based test time (time in seconds each learner spent on reading statements and solving all five test tasks), \( M = 476.75, SD = 313.75 \); rapid computer-based test time (time each learner spent on reading statements and verifying suggested steps for all five test tasks), \( M = 147.97, SD = 20.33 \); test scores for the traditional test, \( M = 13.12, SD = 7.12 \) (58% correct, actual range of test performance scores from 0 to 25); test scores for the rapid test, \( M = 46.19, SD = 11.32 \) (62% correct, actual range of test performance scores from 22 to 67); difficulty ratings averaged over all five tasks for the traditional test, \( M = 5.54, SD = 2.00 \) (actual range of ratings from 1.4 to 9.0; the ratings 1 corresponded to extremely easy tasks); and average difficulty ratings for the rapid test, \( M = 4.68, SD = 1.82 \) (actual range of ratings from 1.0 to 8.6). Correlations between all variables are presented in Table 1.

A Pearson product–moment correlation, \( r(31) = .71, p < .01 \), between scores for the traditional and rapid tests was obtained, with 95% confidence interval extending from 0.48 to 0.85, suggesting a high degree of the concurrent validity for the rapid test scores. A Pearson product–moment correlation between average ratings of task difficulty for the traditional and rapid tests was \( r(31) = .67, p < .01 \), with 95% confidence interval extending from 0.42 to 0.83. There was a significant difference between test times for the traditional and rapid tests, \( r(31) = 5.90, p < .01 \). Test time for the rapid method was reduced by a factor of 3.22 in comparison with the time for the traditional test. The average response time (in seconds) for verification subtasks was \( M = 3.04, SD = 0.81 \), indicating that students actually responded rapidly, as they had been instructed and coached prior to the test. The estimates of Cronbach’s coefficient alpha (.97 for the rapid test and .81 for the traditional test) provided evidence of internal consistency of the scores. The high values are not surprising considering that the tasks measured closely associated attributes (a set of highly related procedural skills in solving a specific class of tasks).

As expected, difficulty ratings in both tests indicated significant negative correlations with test scores. These correlations reflect the fact that more knowledgeable learners experience lower WM load and perceive the tasks as being easier. Test times in the rapid test did not correlate with test times in the traditional test, thus indicating a qualitatively different nature of the rapid test format. The rapid test does not appear to be just an abbreviated version of the traditional diagnostic test. Nevertheless, both tests measure the same construct as indicated by the highly significant correlation between their scores. The rapid test scores did not correlate significantly with rapid test times, indicating that higher scores were likely due to immediately available knowledge base rather than longer time spent on searching for the correct response.

A negative correlation between test times and scores was expected for the traditional test: More knowledgeable learners should obviously finish earlier than less experienced students. Contrary to the expectation, a positive (although nonsignificant) correlation of .19 was obtained. The examination of video records revealed that in kinematics, the computationally intensive and very time-consuming last stage of the solution was rarely reached (or almost immediately abandoned) by novices; however, it was painstakingly performed by more experienced students, thus accounting for extended solution time for these participants.

The results of this experiment indicate highly significant correlations between learners’ performance scores and difficulty ratings on the rapid verification tasks and traditional measures of learners’ task-specific expertise, thus demonstrating a high degree of concurrent validity of the scores obtained from the suggested method. The rapid test was also completed much quicker than the traditional test. To further validate the rapid verification method, I applied it to a different task area in the following experiment.

Experiment 2

This experiment was designed as a correlation study in the task domain of transforming graphs of linear and quadratic functions in mathematics. Two tasks asked students to transform a provided graph of the basic line \( y = x \) into graphs of more complex lines, \( y = -2x + 3 \) and \( y = \frac{1}{2}x - 2 \) (see Figure 2 for an example of a task’s statement). The following two tasks asked students to transform a provided graph of the basic line \( y = x^2 \) into graphs of more complex quadratic functions, \( y = -\frac{1}{3}x^2 \) and \( y = 2(x - 2)^2 \). The tasks required application of two or three of the following operations: flipping a graph because of the minus sign in front of \( x \) or \( x^2 \) (the negative slope), squeezing (expanding) a graph toward (from) the \( y \)-axis according to the value of a coefficient in front of \( x \) or \( x^2 \) (more or less than 1), and horizontal/vertical shifting. The aim of this study was to demonstrate that the rapid verification test in this task domain could be completed rapidly but that the resulting scores would have a high degree of concurrent validity.

Method

Participants

Thirty-three university students who took part in the previous study participated in this experiment. Because the task domain in

Table 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Rapid test scores</td>
<td>-.71**</td>
<td>-.40*</td>
<td>-.53*</td>
<td>-.08</td>
<td>.34</td>
<td></td>
</tr>
<tr>
<td>2. Traditional test scores</td>
<td>-.48**</td>
<td>-.79**</td>
<td>-.11</td>
<td>.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Rapid test difficulty</td>
<td>-.67**</td>
<td>-.24</td>
<td>-.33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Traditional test difficulty</td>
<td></td>
<td>.17</td>
<td>-.21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Rapid test time</td>
<td></td>
<td>-.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Traditional test time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p < .05. **p < .01.
This study was completely different and unrelated to the previous experiment, using the same students could possibly not bias the results. On the contrary, it was expected that the participants’ familiarity with the general format and demands of the rapid testing would eliminate its procedural novelty as a possible negative factor and increase the robustness and validity of results. In addition, the need for pretraining on the rapid testing procedure was also eliminated. In this sense, Experiment 1 could be regarded as a preliminary study in the different task domain, one of the purposes of which was to provide students with extensive pretraining in the rapid testing procedure in realistic conditions.

All students had been taught elementary mathematics in their high school courses. Such courses always include graphs of linear and quadratic functions. Some participants (especially those enrolled in nontechnical university courses) did not deal regularly with such problems and were relative novices in the task domain. Other participants (notably those studying mathematics and engineering courses) were more experienced in the domain. According to the values of total rapid test scores (the maximum possible score of 16), 3 participants were in the range from 0 to 4 (the bottom quarter), 14 participants were in the range from 5 to 8 (the second quarter), 14 participants were in the range from 9 to 12 (the third quarter), and 2 participants were in the range from 13 to 16 (the upper quarter). Students with more expected experience in the domain performed better on the experimental tasks (e.g., students studying mathematics and technical subjects generally scored higher than others in both tests). All participants had not previously encountered tasks formulated using the current format.

Materials and Procedure

The experimental procedure was identical to that used in the previous experiment and included a rapid computer-based diagnostic test and a paper-based test with recordings of students’ on-paper actions and concurrent verbal reports. The sequence of the test administration was counterbalanced.

Rapid diagnostic test. Each task statement was presented for 10 s (this time was established in preexperimental trials as sufficient for comprehending very brief task statements) and was followed by four suggested solution steps for rapid verification. Students were instructed that most of the suggested steps were possible intermediate moves on a way to the solution, but some suggestions could also indicate final answers. Figure 3 shows an example of an incorrect intermediate step for the task represented in Figure 2. Some verification subtasks indicated results of the application of only one operation, whereas other subtasks indicated results of the application of several operations (e.g., flipping and expanding in Figure 3).

Because of the nature of the task, the scoring procedure in this study was different from the cumulative scoring approach used in the previous experiment. For the vector addition tasks, verification subtasks for each solution stage, except the final numerical answer, showed explicitly the fixed cumulative sequence of all prior steps that students would normally perform. For example, a diagram representing the graphical addition of vectors would necessarily show the vectors themselves with assigned numerical values (the essential attributes of the previous solution steps). In contrast, for the tasks used in this experiment, verification subtasks showed only results of the application of an ordered set of possible prior steps. An individual student might not necessarily solve the task using this specific sequence of steps. For example, when constructing a graph of the line \( y = -2x + 3 \), one person could first flip the line \( y = x \), then squeeze the flipped line, followed by the shifting of the squeezed line up. Another individual would prefer to squeeze the original line first, then shift the squeezed line, and finish by flipping the shifted line. Yet another student would shift first and then perform one of the two possible sequences of the remaining steps.

As a result, in this task domain, the verification process is likely to be performed by locating a feature that would immediately

Figure 3. Snapshot of the rapid response window for a graph transformation task.
exclude the suggested step from a list of possible correct steps rather than by comparing the suggested step with different progressive stages of a mentally constructed solution sequence. For example, noticing that a flipped line is depicted for a function with a positive slope, or that an expanded line is depicted when the squeezing operation is required, or that a shift is made in the wrong direction would immediately flag an incorrect step. Because locating a single incorrect operation could be sufficient for the verification purpose, the scoring procedure in this task domain allocated a score of 1 for each correctly performed verification subtask. Therefore, the maximum possible score in the rapid test (4 tasks with 4 verification subtasks for each task) was 4 × 4 = 16. Each diagnostic task was followed by a 9-point subjective rating scale of task difficulty. Students’ performance scores, difficulty ratings, and test times were automatically recorded by the software.

**Paper-based test.** Task statements presented to learners were similar to those used in the rapid test (numerical values in the task statements were changed). Students were asked to provide a complete solution for each task and think aloud while solving the task. A blank coordinate plane was provided on each of the four pages underneath the task statement. Students’ problem-solving steps on paper and their concurrent verbalizations were recorded using a digital video camera. Each diagnostic task was followed by a paper version of the subjective rating scale. A student’s performance on each task was scored as the number of correct steps that the student completed continuously within a short period of time (10–20 s) of starting the solution, based on the analysis of both visual and audio recordings of her or his actions. A total score out of 9 was allocated to each participant (for one task, the range of scores was from 0 to 3; for the remaining three tasks, the ranges were from 0 to 2). The adequacy of this scoring method was backed by clear associations between the scores and expected levels of expertise based on participants’ academic specializations (students studying nontechnical subjects generally scored lower than those studying technical domains). For example, all 12 participants representing technical subjects generally scored lower than those studying technical subjects. Experimental data obtained from university students in task domains of vector addition problems in kinematics and graph transformation tasks in mathematics indicated significant correlations (correspondingly, .71 and .75) between the scores and expected levels of expertise based on participants’ academic specializations (students studying nontechnical subjects generally scored lower than those studying technical domains).

### Results and Discussion

The variables under analysis were paper-based test time (time in seconds each learner spent on solving all four test tasks), $M = 371.27$, $SD = 193.71$; rapid computer-based test time (time each learner spent on solving all four test tasks), $M = 105.52$, $SD = 19.99$; test scores for the traditional test, $M = 3.30$, $SD = 2.85$ (37% correct, actual range of test performance scores from 0 to 9); test scores for the rapid test, $M = 8.33$, $SD = 2.77$ (52% correct, actual range of test performance scores from 2 to 15); average difficulty ratings for the traditional test, $M = 3.63$, $SD = 1.82$ (actual range of ratings from 1 to 7); and average difficulty ratings for the rapid test, $M = 4.05$, $SD = 1.83$ (actual range of ratings from 1 to 7.75). Correlations between all variables are presented in Table 2.

A Pearson product–moment correlation, $r(32) = .75$, $p < .01$, between scores for the traditional and rapid tests was obtained, with 95% confidence interval extending from 0.55 to 0.87, suggesting a high degree of the concurrent validity for the rapid test scores. A Pearson product–moment correlation between average ratings of task difficulty for the traditional and rapid tests was $r(32) = .82$, $p < .01$, with 95% confidence interval extending from 0.66 to 0.91. There was a significant difference between test times for the traditional and rapid tests, $t(32) = 7.97$, $p < .01$. Test time for the rapid method was reduced by a factor of 3.52 in comparison with the time for the traditional test. It should be noted that when the results of the rapid test in this experiment were rescored using a cumulative scoring procedure, similar to that used in Experiment 1, a nonsignificant Pearson product–moment correlation, $r(32) = .21$, $p = .25$, was obtained between scores for the traditional and rapid tests. This significantly lower correlation value than that obtained using a simple scoring method indicates that the selection of an adequate scoring method is essential.

As in Experiment 1, there were significant negative correlations between difficulty ratings and test scores in both tests. These correlations reflect decreased WM load with increased levels of learner expertise. Test times in the rapid test did not correlate with test times in the traditional test, thus indicating a different nature of the rapid test format. As expected, rapid test scores did not correlate with rapid test times, and traditional test scores negatively correlated with traditional test times (less knowledgeable learners required more time to complete the test). The estimates of Cronbach’s coefficient alpha (.96 for the rapid test scores and .85 for the traditional test scores) provided evidence that the reliability (internal consistency) of the test scores was very high.

Thus, significant correlations between performance scores as well as difficulty ratings obtained using rapid test tasks and traditional measures of expertise in this experiment demonstrated a high degree of concurrent validity of the rapid verification scores. The rapid test was also completed significantly quicker than the traditional test.

### General Discussion

Rapidly measuring levels of learner task-specific expertise is required for adapting instructional techniques and formats to changing levels of learner proficiency dynamically, in real time. The experiments described in this article were aimed at validating a rapid verification method for diagnosing levels of learner task-specific expertise. Students were presented with a series of possible (correct and incorrect) intermediate steps of a task solution and were asked to rapidly verify each step (i.e., to establish whether the suggested step was correct). Two experiments were designed to evaluate the concurrent validity of the rapid verification scores in two different task domains. Experimental data obtained from university students in task domains of vector addition problems in kinematics and graph transformation tasks in mathematics indicated significant correlations (correspondingly, .71 and .75) be-

### Table 2. Correlations for Variables in Experiment 2

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Rapid test scores</td>
<td>- .75**</td>
<td>- .55**</td>
<td>- .53**</td>
<td>- .23</td>
<td>- .40**</td>
<td></td>
</tr>
<tr>
<td>2. Traditional test scores</td>
<td>- .39</td>
<td>- .47**</td>
<td>- .33</td>
<td>- .45**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Rapid test difficulty</td>
<td>- .82**</td>
<td>.41**</td>
<td>.12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Traditional test difficulty</td>
<td>- .43*</td>
<td>.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Rapid test time</td>
<td>- .16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Traditional test time</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*p < .05. **p < .01.
tween performance scores on rapid verification tests and measures of expertise that were based on observations of students' detailed solutions of similar tasks. Rapid test times were reduced by factors of 3.2 (for kinematics tasks) and 3.5 (for mathematics tasks) in comparison with traditional test times. Thus, the results indicate that suggested tests could be completed rapidly and that their scores have a high degree of concurrent validity.

Values of Cronbach's coefficient alpha (.97 and .96) provided evidence that the reliability (internal consistency) of the rapid test scores in both experiments was sufficiently high. Task difficulty ratings and test times were used to obtain additional indications of validity of the rapid tests. As expected, task difficulty ratings correlated negatively with test scores in both experiments: Tasks were relatively easier for more expert learners (as measured by the suggested rapid test) than for less experienced participants. In rapid tests in both experiments, there were no correlations between test times and scores, indicating that participants' responses were likely based on their immediately available knowledge structures rather than on time-dependent search processes.

In both experiments, test times in the rapid tests did not correlate with test times in the corresponding traditional tests. This result provides an additional indication that rapid verification tests represent a qualitatively different diagnostic assessment format and not just abridged versions of the corresponding traditional tests. It could be assumed that rapid verification tasks involved a different set of cognitive processes than traditional tests (e.g., searching for a suggested verification solution step in the mentally constructed solution procedure for a task or searching for a feature that would immediately exclude a suggested step from possible correct steps).

Participants in this study ranged from undergraduate to postgraduate university students. As noted by one of the reviewers, the assumption was that each student can be ranked with respect to expertise. There could be a potential confound due to different student abilities: Individuals who finish a graduate school may be brighter than undergraduates in general. However, according to one of the cognitive sciences most important and established results, the level of expertise associated with the available knowledge base is the most important factor influencing student learning and performance, and in most situations, it would override other relevant factors.

Thus, in response to the three research questions formulated for this empirical study, it is possible to state that (a) the suggested online verification procedure is a valid diagnostic method capable of detecting different levels of acquisition of domain-specific knowledge structures, (b) the suggested online verification procedure is a rapid method that could be completed fast enough for real-time applications, and (c) the suggested rapid verification procedure is likely to generalize to different task domains. The rapid verification method, as well as the previously investigated first-step method (Kalyuga & Sweller, 2004), are capable of providing evidence about learner levels of expertise that is diagnostically more valuable and powerful than that obtained in traditional (e.g., multiple-choice) tests. The value of this evidence may approach that of verbal reports, however, with testing times significantly reduced compared with alternative assessment methods.

Applying the rapid verification method to a task generally includes the following steps: (a) establish a sequence of possible main intermediate stages in the solution procedure for the task; (b) for each stage, select representative (either correct or incorrect) solution steps; (c) present the original task statement to a learner for a limited time sufficient to read and understand the statement; (d) present a series of the selected intermediate solution steps to learners, one at a time, with the requirement to quickly verify whether each of the suggested steps could lead to a complete solution of the task; (e) select a scoring method depending on the uniqueness of the solution sequence for the task. If there is only one possible solution path, use a cumulative scoring method (allocate higher scores for more advanced steps that are verified correctly, with one unit added to the score for each level of advancement). If there is a range of possible solution paths, use a simple scoring method (a score of 1 is assigned for each correctly verified step).

For example, in the class of simple linear equations of the type \(ax + b = c\) (solve for \(x\), two main solution stages are subtracting \(b\) from both sides of the equation and dividing both sides of the resulting equation by \(a\). Therefore, for the representative task \(3x + 2 = 5\) (solve for \(x\)), the first step in the design of the rapid verification test is establishing sequential stages of the solution procedure: \(3x + 2 = 5 - 2\); \(3x = 3\); \(3x/3 = 3/3\); and \(x = 1\). At the second step, representative solution steps could be selected, for example, \(3x + 2 - 5 = 5 - 5\) (incorrect); \(3x = 3\) (correct); \(3x/5 = 3/5\) (incorrect); \(x = 2\) (incorrect). Then, the original task statement should be presented to a learner for a few seconds followed by the selected representative solution steps, one at a time. The learner should quickly determine whether the suggested steps could lead to a complete solution of the task. Finally, because there is only one possible solution path in this case, a cumulative scoring procedure needs to be applied. Scores of 4 or 3 should be allocated, respectively, for verifying rapidly the suggested final answer or the step immediately preceding it (in the provided example, \(x = 2\) and \(3x/5 = 3/5\), respectively). A score of 2 should be allocated for a correct answer at the stage of completed application of the first procedural step \((3x = 3)\) and a score of 1 for verifying an incomplete intermediate step in the application of the first step \((3x + 2 - 5 = 5 - 5)\).

This study has been limited to two task classes associated with sufficiently well-structured tasks and predictable solution paths. In such areas, different levels of expert behavior could be described as sequential solution stages a person can complete immediately on starting the solution process. More ambiguous and poorly specified task domains, for example, areas that involve problems with multiple possible routes to solutions, may require special content validation procedures in order to establish expertlike solution paths. The rapid verification method could possibly be used in such relatively poorly structured task domains. Only a limited number of situations or steps representing different possible stages of valid solution procedures could be selected and included into rapid verification subtasks. For example, for a medical diagnosis task, a sequence of progressively more advanced stages of testing different hypotheses (including both suitable and inappropriate steps) could be presented for rapid verification. Because of the variety of possible solution paths in this case, a simple scoring procedure should be applied. In further research, the general utility and limits of usability of the method, especially in poorly structured task domains, need to be established.

In some areas, the rapid assessment approach (both the first-step and rapid verification methods) could be more suitable for measuring level of expertise of relatively advanced learners rather than
for thorough cognitive diagnosis of novices. Novice learners may have knowledge deficits of types that could not be anticipated in advance when selecting relevant possible solution steps for verification or programming the scoring engine (e.g., linguistic comprehension problems, insufficient factual knowledge, lack of basic metacognitive planning and monitoring skills). Most of these types of knowledge are usually taken for granted when dealing with relatively more experienced learners.

As a form of dynamic assessment, the method could be instrumental in developing adaptive expertise (Bransford et al., 1999) by enabling the selection of optimal environments for building flexible skills. On the basis of the fine-grained diagnosis of the acquired knowledge and skills in a task domain, such environments would provide learners with individually tailored instructional support that is optimal for moving on to the next level of expert performance. Acquisition of task-specific expertise is necessary for releasing cognitive resources for dealing with more nonroutine and creative aspects of expert performance. With learner-tailored, just-in-time forms of support, learners would be capable of efficiently handling relatively new situations without a cognitive overload and associated loss of interest in the task.

A preliminary study indicated that the suggested diagnostic method could be successfully used for the dynamic selection of appropriate levels of instructional guidance that are optimal for learners with different levels of task-specific expertise (Kalyuga, 2006a). The method was used in an adaptive tutorial in the domain of kinematics for online monitoring of current levels of learner expertise. At the beginning of a tutoring session, each student was allocated to an appropriate level of guidance according to the outcome of the initial rapid verification test. Depending on the outcomes of rapid verification probes during instruction, the learner was allowed to proceed to the next stage or required to repeat the same stage and then take the rapid test again. At each subsequent stage, a lower level of instructional guidance was provided to learners by gradually fading fully worked-out steps and increasing the number of steps that learners had to complete on their own. Also, a higher level of the rapid verification task was used at the end of the stage. The adaptive approach proved to be superior to traditional nonadapted instructional formats. In the future, more comprehensive studies are needed for comparing superior to traditional nonadapted instructional formats. In the next level of expert performance, acquisition of task-specific expertise is necessary for releasing cognitive resources for dealing with more nonroutine and creative aspects of expert performance. With learner-tailored, just-in-time forms of support, learners would be capable of efficiently handling relatively new situations without a cognitive overload and associated loss of interest in the task.

References


Received August 24, 2006
Revision received August 9, 2007
Accepted August 10, 2007