

**ON-LINE ASSESSMENT OF HIGHER-ORDER THINKING SKILLS:
A JAVA-BASED EXTENSION TO CLOSED-FORM TESTING ®**

David A. Thomas, Giray Ökten, and Paul Buis
Ball State University
USA

WWW-based mathematics and statistics courses frequently incorporate machine-scorable items (i.e., True-False, Multiple Choice, and Matching) in both formative and summative assessments. For instance, WebCT and BlackBoard provide interfaces for the development and delivery of closed-form quizzes and examinations. Using these technologies, it is relatively easy to determine whether students possess detailed factual knowledge. It is much more difficult, using these technologies, to assess higher order thinking skills. This paper presents a Java-based extension to closed-form testing that may be better suited to assessing higher-order thinking skills.

INTRODUCTION

For over a decade, reform movements in mathematics and statistics education have emphasized teaching for understanding. One outcome of this activity has been a renewed interest in the assessment of higher-order thinking skills. We are particularly interested in assessment procedures and technologies that reveal more about what students know, as opposed to what they don't know, and how students think about mathematical and scientific ideas. Our current research focuses on the development of strategies and technologies for assessing higher-order thinking skills in WWW-based mathematics and statistics courses. We believe that, until achievement in WWW-based courses is assessed as rigorously and reliably as in on-campus courses, WWW-based courses will not be credible with or acceptable to many institutions and individuals. This paper describes the development and testing of a new, Java-based, closed-form assessment tool that addresses that goal. We call the item type that is the basis of our approach a *grid*.

SELECTION TASKS

In grid-based assessment, information (e.g., text, formulas, figures, numbers) is presented in an array of cells called a grid. A task or set of tasks is presented apart from the grid. In composing a response, the student scans the grid, looking for the cell or cells that collectively constitute a correct and complete response to the task. Depending on the task, a correct answer may consist of one or more cells. Figure 1 illustrates a type of grid-based task called a *selection*.



Task #1 Which cells contain equivalent fractions, decimals, percents, or shaded portions of the whole?	CELL 1 66%	CELL 2 	CELL 3 1.5
	CELL 4 .40%	CELL 5 .04	CELL 6 2/5
	CELL 7 4/6	CELL 8 	CELL 9 .40

Figure 1. Sample Selection Task.

Selection tasks ask the student to identify sets of grid cells that stand in a particular relationship defined by the task. For example, the contents of certain grid cells may be related in that they are numerically or algebraically equivalent, descriptive or illustrative of a particular

concept or process, recognizable parts of some whole, or steps in an algorithm. Such associations are the very stuff of declarative knowledge structures and are the basis of conceptual knowledge in mathematics and statistics. In the case of Task #1, cells 2, 6, and 9 contain equivalent representations of the same number. Therefore, a correct response to Task #1 is the set {2, 6, 9}.

Figure 2 shows a planning table used by the authors to design selection Task #1. Planning tables of this sort are never shown to students, for they reveal both the solution to the task and traps set to detect specific misconceptions. The left side of the table shows all pair-wise cell relationships in a correct response. The right side shows pair-wise cell relationships associated with misconceptions set as error traps by the teacher.

	Equivalent Forms (Cells)				Error Traps (Cells)			
	fractions	decimals	percents	graphics	fractions	decimals	percents	graphics
fractions								
decimals	6&9				5&6			
percents					1&7	4&5		
graphics	2&6	2&9			2&7 6&8	3&8	2&4	

Figure 2. Planning Table.

Selection tasks are scored using a computerized matrix algebra procedure that compares all of the pair-wise cell relationships in a student response to the pair-wise cell relationships in a correct response. This comparison is used to generate an item score (i.e., partial credit) for the task. The same computer program that scores student responses is used to generate formative feedback messages keyed to "trapped" misconceptions. For instance, students who have difficulty converting fractions to decimals might believe that cells 5 and 6 are equivalent. In a fully developed grid-based assessment system, students will receive feedback focused on correcting the specific misconceptions implicit in their responses. At the ICOTS-6 conference, examples of this sort of scoring and feedback will be demonstrated as opportunity permits.

SEQUENCING TASKS

Procedural knowledge is an important aspect of many disciplines. In particular, knowing "what comes next" is critical in many scientific, mathematical, and statistical contexts. A *sequencing* task requires the student to order all of the cells in a grid according to some criterion. For instance, given a grid containing

- Names of selected chemical elements, science students might be asked to order the elements on the basis of their atomic numbers;
- A mixed set of integers, fractions, and decimals, mathematics students might be asked to order them smallest to largest;
- Steps in a hypothesis testing procedure, statistics students might be asked to order the steps.

Like selection tasks, sequencing tasks are scored using a procedure that compares all of the pair-wise cell relationships in a student response to the pair-wise cell relationships in a correct response. This approach avoids a distracting and confusing issue: The enormous number ($n!$) of permutations possible in a set of n cells. Instead, our scoring scheme focuses on characterizing the internal "orderliness" of a sequence by inspecting only $(n^2 - n)/2$ pair-wise cell comparisons. These inspections are easily handled using simple computational matrix algebra.

COMBINATION TASKS

Combination tasks require students to first select and then sequence a subset of grid cells. For instance, the superintendent of a power plant might want to know whether the plant's operating engineers can reliably and consistently select a suitable sequence of actions from a complex set of possible responses when presented with various crises. A less dramatic circumstance is shown in Task #2, where statistics students are asked to select then sequence a set of actions in order to decide which of two standardized test scores represents a superior

performance. The format of the cell contents indicates that two sets of calculations are required, followed by a decision. The correct response to this task is the ordered sequence 1-6-4-8.

Task #2
 You wish to compare scores from two standardized tests to determine which score is better. Scores from both tests are normally distributed. List the cell numbers that contain the proper steps in the order that you would perform them.

1 Raw score(s) minus mean score(s)	2 Divide by mean(s)	3 Compare raw scores
4 Compare z-scores	5 Mean score(s) minus raw score(s)	6 Divide by standard deviation(s)
7 Divide by raw score(s)	8 Select larger z-score	9 Select smaller z-score

Figure 3. Combined Selection & Sequencing Task.

PRELIMINARY RESULTS

We have developed methods for aggregating scores on individual selection, sequencing, and combination items to yield an overall examination score. We have also studied the expected outcome of random guessing relative to each item type and their aggregations. In the course of these investigations, we have developed closed-form descriptions of the expected outcomes that agree closely with simulations created using independent methods. On the basis of these findings, we believe that grid-based assessment may provide a "finer grained picture" of student knowledge structures and misconceptions than traditional closed-form assessment.

We have also conducted a limited number of informal investigations to determine whether students perform differently on grid-based items than they do on comparable open-ended items. For example, a group of 30 students enrolled in the course Mathematics for Elementary Teachers at Ball State University were given a paper-and-pencil statistics quiz that included Task #2 above and a related open-ended item. Full or partial credit was assigned to each response. When these scores were analysed, it was clear that differences in performance on the two item types were no greater than one might expect between two open-ended items of the same type ($r = 0.82$). That is, students who did well on traditional open-ended items, also did well on grid-based items, and so on. While a great deal of work remains to be done on the issues of validity and reliability of measurement using grid-based methods, we are encouraged to note that average students appear to adapt quickly to the new item types. Consequently, we are currently planning additional trials addressing a wide range of technical issues and content areas.

COMPUTER IMPLEMENTATION

Our first step in developing a computer environment for grid-based testing was to create a spreadsheet model of the scoring procedure. In the spreadsheet, we were able to view all of the variables and algorithms employed in the model simultaneously. This made it easy to "track" the manner in which the model treated student responses and to verify that the model was performing as intended. While a full discussion of the scoring procedures is beyond the scope of this paper, the fundamental concept may be illustrated as follows in the case of a selection task. Related strategies have been developed for analysing responses to sequencing and combined tasks.

1. Given a selection task with n cells and a planning table like that shown in Figure 2, the instructor creates an $n \times n$ matrix, C , called the correct response matrix. In this matrix, each entry C_{ij} is either a 1 or a 0, depending on whether the contents of cells i and j do or do not meet the selection criterion of the task. Matrix C characterizes all pair-wise cell relationships in a correct response.
2. Using the same procedure, a similar matrix R is created that characterizes a student's response to the item. Matrix R characterizes all pair-wise cell relationships in the student's response.
3. When matrix R is subtracted from matrix C , an error matrix E is obtained like that shown in Figure 4. In this matrix, a 1 indicates an error of omission, i.e., an essential pair-wise

relationship is missing. A -1 indicates an inclusion error, i.e., a false relationship is included in the response. And a 0 indicates a correct response. Because of the symmetry of this matrix, only the lower triangular portion need be considered.

4. The overall score for the item is computed as the number of 0's in this lower triangular matrix divided by the number of elements in that portion of the error matrix. For instance, the response {5, 6, 9} to Task #1 produces the error matrix seen in Figure 4. Reports of this sort are meant for the instructor, not the student.

	CELL 1	CELL 2	CELL 3	CELL 4	CELL 5	CELL 6	CELL 7	CELL 8
CELL 1								
CELL 2	0							
CELL 3	0	0						
CELL 4	0	0	0					
CELL 5	0	0	0	0				
CELL 6	0	1	0	0	-1			
CELL 7	0	0	0	0	0	0		
CELL 8	0	0	0	0	0	0	0	
CELL 9	0	1	0	0	-1	0	0	0

Legend	-1	Inclusion Error (An incorrect association is included in the response.)
	0	Correct
	1	Omission Error (A necessary association is missing in the response.)
Score:	88.9%	Of the indicated pair-wise associations are correct (0: 32/36)
	5.6%	Inclusion Errors (-1: 2/36)
	5.6%	Omission Errors (1: 2/36)

Figure 4. Error Matrix.

In order to deploy grid-based assessment across TCP-IP networks, we are currently using Java to implement scoring and reporting procedures. At the moment, we have an alpha level version of the presentation and analysis modules in testing. At ICOTS-6, we hope to also have alpha level versions of a task editor available for demonstration. Our goal is to make this tool available to interested faculty at Ball State University sometime in 2003 for trials in undergraduate mathematics and statistics courses.

DISCUSSION

Our interest in grid-based assessment was first motivated by Johnstone's (1987, 1979) work with paper-and-pencil grids in university chemistry courses. What we have added to his work are powerful response analysis and feedback models and computer-based implementations in the context of teaching mathematics and statistics. We are also conducting human-factors studies on the reliability and validity of measurement in grid-based assessment. From preliminary studies, it is apparent that these measures are likely to depend heavily on the quality of the individual tasks used in any given assessment. Since this is also true of conventional closed-form testing, we hope to identify practices that contribute to the quality of conventional tasks that are transferable to grid-assessment tasks. We believe that a critical factor in the development of grid-based assessment will be the creation and testing of mathematical and statistical assessments by independent scholars. ICOTS-6 participants interested in participating in our research in this manner are invited to contact David Thomas via email at dthomas@bsu.edu. Full contact information is available at the URL <http://www.cs.bsu.edu/homepages/dathomas/>.

REFERENCES

- Johnstone, A.H. (1987). Methods of assessment using grids. Talk given at the Science-Math Education Centre at Curtin University of Technology, Perth, Western Australia.
- Johnstone, A.H., & Mughol, A. (1979). Testing for understanding. *School Science Review*, 61 (214), 147-150.