Automated educational / academic skills screening: using technology to avoid or minimise effects of more formal assessment

John Izard
School of Education
RMIT University
john.izard@rmit.edu.au

and

Elspeth McKay
School of Business Information Technology
RMIT University
elspeth@rmit.edu.au

The focus of this study is the provision of enhanced opportunities for returning to study or vocational training for adolescents and young adults (aged from about 15 to 25 years) after experiencing a mental health episode, and the monitoring of their educational progress during rehabilitation. Both the young people and their support workers need to establish current educational / academic performance in order to make sensible choices about future study or employment. But formal assessment techniques are associated with high levels of anxiety: it was vital to devise another way of collecting this information. This project provides an innovative use of information communications technology (ICT) services to assess the educational achievement level of this at-risk learner group. The paper reports on the design, development, implementation and evaluation of the pilot system for assessing young peoples' potential to participate in appropriate educational programmes.

Context for the Study

Those who work with young people, aged from about 15 to 25 years, cannot assume that the prior level of educational achievement is still an appropriate description of current level after a mental health episode. Choosing an appropriate point of entry for study or vocational training *without* assessing current level of achievement would be foolhardy. However, these young people often find the usual test instruments and examinations distasteful and anxiety-producing. Many young people in this client group have not been involved in a formal learning environment for some time, some as long as 10 years. The formal approaches to assessment bring back bad memories of their mental health episode or its precursors, so they resist such approaches (Ring, 2003, personal communication). Furthermore, many of these young people have difficulty in selecting an appropriate path without skilled help, either in education or employment, and become frustrated and disappointed when they fail at tasks for which they are not suited currently.

But choosing an appropriate point of entry for study or vocational training requires knowledge about a person's capabilities in terms of intellectual skills, cognitive strategies, verbal information, motor skills, and attitudes (Gagne, 1985). The intellectual skills capabilities relate to how an individual interacts with the environment, deals with procedures ranging from translating simple instructions to dealing with more complex procedures, symbols and algorithms including distinguishing hierarchical relationships necessary for problem solving and organization. This procedural knowledge is sometimes described as *knowing how*. The cognitive strategies include the ways people choose to make decisions, how they remember, and how they incorporate experience from past learning events. Verbal information impacts on these capabilities through declarative knowledge,

sometimes described as *knowing that*. The capacity to execute movements (motor skills) contributes to these capabilities through writing, assembling, constructing and refining. Attitudes influence an individual's choice of particular personal actions such as deciding whether to study and which-topic to study next.

This paper describes a pilot study to design and develop an automated educational / academic skills screening that used technology to avoid or minimise some of the more formal features of assessment. The Educational/Academic (skills) Screening for the Young pilot system (known as EASY) was developed and evaluated. The project was a Telematics Trust funded collaborative project between members of the Department of Education and Training at ORYGEN Youth Health and RMIT University, seeking to model a real-world system in situ to reflect the special needs of young people who are dealing with the residual effects of recovering from a mental illness. The project team sought to increase the self esteem of young people in need of assistance to achieve personal goals and improve life-long learning characteristics; by building an innovative system that was interesting to young people, fun and easy to use, and that delivered positive output for users.

It was hoped that this screening tool would assist participants and carers in selecting more appropriate endeavors, leading to a greater probability of success and therefore re-building their self esteem. It was recognised that the steady appropriate development of participant knowledge and skills would assist in their successful rehabilitation and entry into the workforce, thereby reducing their need for health services and welfare support from taxpayer funding. The authors believed that the project could show how implementation of technology could enhance the quality of life for young people, and that the knowledge gained from this project would be highly transferable to other areas of youth work in general, and specifically within the context of youth disability.

Plan for the Assessment

The intention of the project team was to design~~,~~ and develop an academic-skills screening tool as a prototype system and test it with young people who are dealing with the residual effects of recovering from a mental illness. We hoped that it would:

- Cover a range of skill levels from novice to experienced,
- Be age-appropriate in content and presentation,
- Be user friendly, not requiring a high degree of information technology skills,
- Have enough functionality to demonstrate the benefits of using such a system.

The results of the screening would be interpreted by support staff in the context of individual goals, past experiences and current presentation. Later we hoped to extend the prototype measurement instrument to provide better estimates of knowledge / performance levels that would help carers identify what was within reach. The longer-term goal of this project was to design and build a fully adaptive Web-enabled skill differentiation system.

The authors were aware that choosing an appropriate point of entry for further study or vocational training required information about the individual's current standing in educational achievement. The project team knew that reported achievement levels become dated due to subsequent experience and learning, and furthermore, that mental health episodes could make the prior reported achievement level obsolete. Although there are instruments that measure academic skill in terms of performance of a reference or norm

group, few differentiate what an individual knows (about something), from what they do not know (about that something). Yet knowing what a person knows or does not know is more powerful in helping that person decide what to tackle next (Izard, 2002). The essentially rank-order information of reference group reporting did not provide adequate evidence of individual skills and knowledge.

The research team conducted a thorough analysis of the knowledge acquisition environment relating to young people wishing to return to study or participate in vocational re-establishment programmes. A cognitive performance capability matrix was developed to fill the gaps found in existing measurement techniques (see Figure 1). This matrix identifies educational/academic tasks in terms of basic to complex, and measurable cognitive performance outcomes in terms of declarative (*knowing that* – verbal information) to procedural knowledge (or *knowing how* – a cognitive strategy).

Accordingly, we focussed on a range of tasks (that looked somewhat different to conventional test items) to provide evidence of

- Literacy skills including comprehension and interpretation
- Numeracy skills including measurement and estimation
- Problem solving and organization
- Memory and concentration levels

The test comprised a wide range of assessment tasks to provide achievement evidence relating to everyday matters. The overall test structure is shown in Table 1. Tasks included instances where the person had to identify key elements from safety symbols, diagrams, pictures, tables, or written passages, answer questions about magnitudes such as length, width, distance and temperature, aspects of food purchasing, safe storage of food, interpreting recipes, and economical purchases. In the initial data collection it was hoped that all persons attempting the test would attempt all items so that these could be checked for validity and calibrated for difficulty.

In the first set of tasks (*Safety*) all items were scored correct/incorrect (1/0). The maximum total score for this set of tasks was 4. The second set of tasks (*Magnitude*) had 14 items, with 11 scored correct/incorrect (1/0) and 3 scored as partial credit items. (Partial credit items assign different marks according to the quality of the student response. For example, item N3 was scored 0, 1, 2, or 3 according to quality of the response. The other partial credit items (L3, M4 and N2) were scored 0, 1, or 2 according to quality of the response.

A comment is required on items G3, G4, and SU. The required response to G3 was an amount, the total cost of three items illustrated. The required response to G4 was an amount, the change from paying for the G3 items with a $20 note. If a person made an error in G3 it was likely that G4 would be wrong too. Item SU gave credit for all cases where the G3 amount plus the G4 amount was $20. There is a risk with these types of items that some of the information gathered will be redundant because the individual items are not independent. (This issue is taken up again later when the analyses are reported below.)

In the third set of tasks (*Food*) all items were scored correct/incorrect (1/0). The fourth set of tasks (*Identification*) had 18 items, with 17 scored correct/incorrect (1/0) and item T1 scored as a partial credit item scored 0, 1, or 2 according to quality of the response.

<table>
<tr><td rowspan="6">Automated Educational/ Academic Skills Evaluation for Young People: (EASY) system Proof-of-Concept</td><td colspan="2">Declarative Knowledge</td><td colspan="3">Procedural Knowledge</td></tr>
<tr><td>Band-A</td><td>Band-B</td><td>Band-C</td><td>Band-D</td><td>Band-E</td></tr>
<tr><td>Verbal information skill</td><td>Intellectual skill</td><td>Intellectual skill</td><td colspan="2">Cognitive strategy</td></tr>
<tr><td>Concrete concept</td><td>Basic rule</td><td>Higher-order-rule</td><td rowspan="2">Identify sub-tasks</td><td>Knowing the “how”</td></tr>
<tr><td>Knows basic terms</td><td>Discriminates</td><td>Problem solving</td><td>Recall simple pre-requisite rules & concepts</td></tr>
<tr><td>Knows “that”</td><td>Understands concepts & principles</td><td>Applies concepts & principles to new situations</td><td>Recognizes unstated assumptions</td><td>Integrates learning from different areas into plan for solving a problem</td></tr>
<tr><td rowspan="2">Learning Domain Task Code = R</td><td rowspan="2">Texts</td><td colspan="4">Aspects of Language</td></tr>
<tr><td>Contextual understanding</td><td colspan="2">Linguistics</td><td>Strategies</td></tr>
<tr><td>Reading (constructing meaning from print & non-print)</td><td>Literature (books, etc)</td><td>Everyday texts (telephone conversations, notices)</td><td colspan="2">Media texts (newspapers, internet, TV, Video, CD-ROM)</td><td>Workplace texts (letters, resumes, reports, etc)</td></tr>
</table>

Figure 1 Measuring academic performance: Cognitive performance capability matrix for instructional objectives (McKay and Izard, 2004)

Table 1- Tasks, Codes, Components and Maximum Scores for Test Components

Task Name	Codes (No. of tasks)	Max. Score
Safety	Q1, Q2, R1, R2 (4)	4
Magnitude	E1, E2, E3, F1, G1, G2, G3, G4, SU*, H1, J1, J2, K1, L1, L2, ***L3***, M1, M2, M3, ***M4***, M5, N1, ***N2***, ***N3*** (24)	29
Food	V1, V2, V3, V4, W1, X1 (6)	6
Identification	A1, B1, B2, C1, C2, C3, D1, D2, D3, ***D4*** S1, ***T1***, T2, T3, T4, T5, T6, U1 (18)	20
Consumer Awareness	Y1, Z1, Z2, Z3, Z4 (5)	5

Note: * indicates an item that used information from two other items in order to give credit for calculation skills even though the one of the responses to other items was incorrect; ***italics*** are used to identify partial credit items.

All items in the fifth set of tasks (*Consumer Awareness*) were scored correct/incorrect (1/0). The possible scores on this test of 57 items ranged from 0 to 63. (A video segment about a job interview provided responses to some open-end questions. These responses will be of considerable assistance in generating an assessment framework for that task.)

Plan for the Computer Presentation

We chose a human-computer interaction (HCI) framework that involved linking people, educational practice, technology and the environment. Choosing the correct model for HCI can be very difficult (Preece, 1994) because of the inter-disciplinary nature of HCI. Systems design is complex, and changing the nature of any component can lead to unexpected outcomes if the complexity is not fully understood during the design phase of system development. The factors that had to be considered are shown in Figure 2.

<table>
<tr><td colspan="2">ORGANIZATION FACTORS
Training, job design, politics, roles, work organization</td><td>ENVIRONMENTAL FACTORS
Noise, heating, lighting, ventilation</td></tr>
<tr><td>HEALTH & SAFETY FACTORS
Stress, headaches, Muscoluo-skeletal disorders</td><td>Cognitive processes & capabilities

THE USER

Motivation, enjoyment, satisfaction, personality, experience level</td><td>COMFORT FACTORS
Seating, equipment layout</td></tr>
<tr><td colspan="3">USER INTERFACE
Input devices, output displays, dialogue structures, use of colour, icons, commands, graphics, natural language, 3-D, user support materials, multi-media</td></tr>
<tr><td colspan="3">TASK INTERFACE
Easy – complex, novel, task allocation, repetitive, monitoring, skills, components</td></tr>
<tr><td colspan="3">CONSTRAINTS
Costs, time scales, budgets, staff, equipment, building structure</td></tr>
<tr><td colspan="3">SYSTEM FUNCTIONALITY
Hardware, software, application</td></tr>
<tr><td colspan="3">PRODUCTIVITY FACTORS
Increase output, increase quality, decrease costs, decrease errors</td></tr>
</table>

Figure 2 Human factors in human-computer interaction (HCI) (adapted from Preece, 1994, p.31)

McKay and Izard (2004) have described the system development process in more detail.

An important feature of this system was to make access to the information easy. The system utilized touch screen technology: the system functioned without a keyboard and mouse. The navigation was kept simple. To exit, the user presses the QUIT button. At this point the particular assessment session is over. Re-entry is possible again by re-pressing one of the buttons on the Home Page. To operate the EASY prototype, users are required to press any button on the EASY Home Page. (see Figure 3) This operation delivers system sub-level activity screens where there are a number of questions. Due to the special design feature for skills assessment, the usual screen chattels for navigation do not display. Instead there are invisible controls that the facilitator presses that include buttons for NEXT screen (top right corner), PREVIOUS screen (top left corner), HOME (bottom right) and EXIT (bottom left).

Figure 3 Prototype Home Page

Figure 4 Sub-level Menu

In keeping with the 5-star principles of instruction (Merrill, 2003) the prototype offered video on demand (VOD) (Okamoto, Matsui, Inoue and Cristea, 2000). These vignettes presented everyday activities related to health and safety issues, organizing your study papers, preparing for an interview, and so on (Figure 4).

Plan for the Study

The study was arranged in phases, as shown in Figure 5.

Phase	Duration	Purposes
1: Needs Analysis	2-months (Jan-Feb. 2003)	Consult with • Young people in the community aged 15-25 years, • learning programme/course developers, • educational programme selection officers, • vocational rehabilitation specialists (including psychologists), • self-help agency staff members Determine anticipated knowledge/skill requirements for various learning events, Devise suitable instrumentation to measure characteristics of the young people's skill acquisition capabilities
2: Develop System Specifications	2-months (March-April, 2003)	Design capability matrix to identify educational/academic tasks (basic to complex; visual [pictorial] and text-based) and measurable cognitive performance outcomes, for the cognitive performance/ability scale Seek and obtain required Ethics Committees' approval (RMIT University and ORYGEN Youth Health)
3: Build System	3-months (May-July, 2003)	Design specially devised assessment tasks (visual and text), Review the EASY content and system, Make essential alterations in consultation with all project team members.
4: Implement System	1-month (Aug, 2003)	Install the EASY system, Conduct a trial with a sample group at the ORYGEN Youth Health facility (young people attending the centre for vocational rehabilitation purposes), Amend content and system as necessary, Devise record protocols for evaluation purposes.
5: Evaluate System Use	2-months (Sept-Oct, 2003)	Conduct trial with expanded sample group at the ORYGEN Youth Health facility (young people attending the centre for vocational rehabilitation purposes), Record reactions of the participants as secure codes, Tabulate all responses for each individual.
6: Conduct Data Analysis	1-month (Nov, 2003)	Using the QUEST test analysis software to estimate levels of achievement and content/skills at each level, Review results with project team.
7: Closure	1-month (Dec, 2003)	Document findings as a Completion Report, Review of report by project team, Distribute report of findings and recommendations for future research to ORYGEN Youth Health, and funding agency (Telematics Course Development Trust), Disseminate through conference presentations and journal publications.

Figure 5 Plan for the Project

Results

The trial of the prototype system was conducted over a three-month period. The sample group of participants was attending a vocational rehabilitation programme. Preliminary feedback, including that of several non-computer users, was very positive. The users found the system easy to use, informative, and could relate to the characters in the various VOD vignettes. The study was carried out mostly as intended. The most significant exception was the design of the specially devised assessment tasks (visual and text) in Phase 3.

A contract had been let to a company to prepare the items. This company was provided with a detailed specification with examples of item types but was not able to deliver more than a modest number of items over a period of three months. The contractors were not able to have the whole test finished within the time frame imposed by the funding arrangements, to rectify some errors in the items supplied, nor to meet the requirements for full-screen presentation on the computer screen. The items supplied could not be used in their current form. Accordingly, one of the project team undertook to develop a new set of the necessary items and to provide the associated photographs in time to administer the prototype system before participants departed at the end of the academic year. This target was achieved: all of the items and photographs were delivered within two weeks.

The tasks were arranged by another contractor as a Macromedia Flash presentation with hidden buttons to allow the administrator of the test to control its administration. This allowed the test administrator for the trials to order the sets of tasks according to her judgments of task difficulty. Table 1 above shows the tasks in the order she chose. Subsequent research will investigate self-administration by the person attempting the test with test-response success rates modifying the next to be items presented. The administrator will retain control over the output data provided.

The analyses required a person-by-item matrix of data. It was proposed that each student would attempt all of the tasks so all tasks could be calibrated against each other. This was necessary to take account of variation in task difficulty, and to allow comparable scaled scores to be obtained in conjunction with new items and to enable appropriate scale scores to be estimated for students who attempt these different tasks in subsequent use.

The first run was to check the technical properties of a test comprising these tasks. It was a check that items fitted the model of increasing competence with increasing score, and to identify any items that should be deleted. The summary results are shown in Table 2. These results show that the items were able to separate the students in a consistent way, and in turn, that the students were separated by the items. The scaled mean difficulty for the items is an artifact of the type of analysis. On the same scale, the students had a similar mean. The range of the items was wide. This is useful in a caring context because the test tasks can assess improvements over time (as shown by an improvement in scaled student score). If there had been a ceiling effect then such a test would have been ineffective for showing progress: those with perfect scores would be likely to achieve perfect scores again after further work but the “evidence” from the testing would imply that there had been no change. Any students with perfect scores are beyond the range of the test and therefore no increase in achievement could be detected.

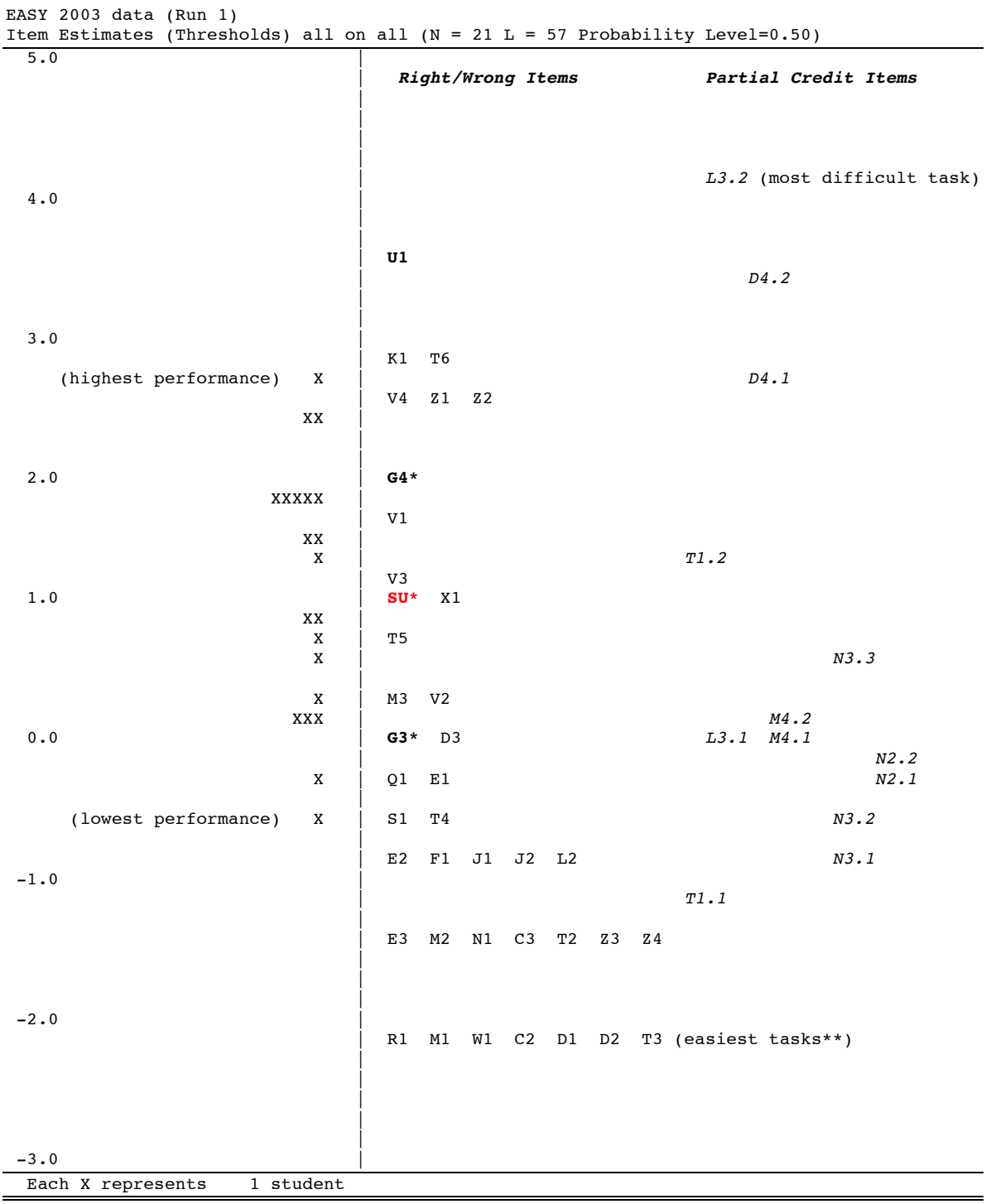
Figure 6 shows all of the tasks on the right-hand side of the variable-map diagram. More difficult tasks are at the top, and the easier tasks are at the bottom. The vertical line separating items from students represents a skills-achievement continuum. Students are shown (anonymously) on the left-hand side of the variable-map diagram, with highest achieving students at the top. (More able students can do more difficult items.) No person had a zero or a perfect score. All persons received a score.

Table 2 Summary Results: Test with *Safety*, *Magnitude*, *Food*, *Identification*, and *Consumer Awareness* Tasks (Run 1)

Item Estimates (Thresholds) all on all (N = 21 L = 57 Probability Level=0.50)	Case Estimates all on all (N = 21 L = 57 Probability Level=0.50)
Summary of item Estimates ========== Mean 0.00 SD 1.66 SD (adjusted) 1.46 Reliability of estimate 0.78	Summary of case Estimates ========== Mean 1.14 SD 0.89 SD (adjusted) 0.80 Reliability of estimate 0.81
Fit Statistics ========== Infit Mean Square Outfit Mean Square Mean 0.99 Mean 1.14 SD 0.15 SD 1.15	Fit Statistics ========== Infit Mean Square Outfit Mean Square Mean 1.00 Mean 1.14 SD 0.33 SD 0.90
Infit t Outfit t Mean 0.04 Mean 0.19 SD 0.55 SD 0.85	Infit t Outfit t Mean 0.01 Mean 0.14 SD 1.06 SD 1.23
0 items with zero scores 12 items with perfect scores	0 cases with zero scores 0 cases with perfect scores

(To place a student on the scale we have to know both what he/she can do and what he/she cannot do. For students with a score of 0, we do not know what they can do. Similarly, for students with perfect scores, we do not know what they cannot do.) Twelve items were discarded from the analysis (but not from the test) because every person was correct on those items. (The items were retained on the test to give respondents encouragement to continue the test.) Detailed records by the test administrator showed that there was a minimal need to explain words used in the test.

Items (task components) are shown on the right-hand side of the diagram (see Figure 6). The first column on the right shows the thresholds of components scored correct/incorrect (1/0) (see Right/Wrong Items). The easiest correct/incorrect items are shown in the bottom part of the diagram and the most difficult are shown in the top part. For example, the threshold for *Magnitude* task J1 is between the scale values of 0.0 and –1.0 but closer to –1.0. Persons above this threshold are more likely than not to be correct on this task. The higher the person is above the item threshold, the more likely the person scored 1 for the item. . If persons are at this threshold they have a 50% probability of scoring 0 and a 50% probability of scoring 1 on this component. If persons are below this threshold they are more likely than not to be incorrect on this component. The lower the person is below the item threshold, the more likely the person scored 0 for the item.



* Note 1: Information from items G3 and G4 contributed to success on item SU in order to give credit for calculation skills in cases where one of the responses to other items was incorrect. *Italics* are used to identify partial credit items.

** Note 2: These were the easiest tasks where items remained in the analysis. Because items Q2, R2, G1, G2, H1, L1, M5, A1, B1, B2, C1 and Y1 had perfect scores those items were not used in the analysis.

Figure 6 Variable Map Showing Right/Wrong and Partial Credit Item Thresholds

Where items are on the same horizontal level, the probability of being correct is the same or similar for each of those items. For example, items E2, F1, J1, J2, and L2 are all 0/1 scored items of similar difficulty. They represent a comparable level of difficulty to a score of 1 on the partial credit item N3 (shown on the variable map as N3.1).

The partial credit items are to the right-hand side of this variable map (Figure 6). Thresholds for partial credit items are interpreted in a similar way. For example, there are two thresholds for *Identification* item T1. The threshold for a score of 1 on item T1 is shown T1.1 and is between the scale values of –1.0 and –2.0. Those above this threshold are more likely than not to score 1 or more on this task. The higher the person is above the item threshold, the more likely the student scored 1 or more for the item. If persons are at this T1.1 threshold they have a 50% probability of scoring 0 and a 50% probability of scoring 1 on this task. If persons are below this threshold they are more likely than not to score 0 on this task. The lower the person is below the item threshold, the more likely the person scored 0 for the item.

The threshold for a score of 2 on item T1 is shown as T1.2 and is between the scale values of +1.0 and +2.0. Those above this threshold are more likely than not to score 2 on this task. The higher the person is above the item threshold, the more likely the person scored 2 for the item. If persons are at this T1.2 threshold they have a 50% probability of scoring 1 and a 50% probability of scoring 2 on this task. If students are below this threshold they are more likely than not to score 1 or 0 on this task. The lower the person is below the item threshold, the more likely the person scored 1 or 0 for the item. It is easier to score 1 on item T1 than to score 2 on the same item. Note that the *Magnitude* item N3 has 3 thresholds: N3.1, N3.2 and N3.3 representing scores of 1, 2 and 3 on N3 respectively. The threshold for scoring 1 on S1 or T4 is at approximately the same level as scoring 2 on N3 (N3.2).

The fit to the test model in effect size terms is summarised in Figure 7. This diagram shows magnitude of effect rather than statistical significance. The dotted lines either side of the vertical line (under the 1.00) show what we decided was the acceptable range of infit mean squares. One item (G4) is shown to the left of the left-hand dotted line. On the basis of actual person responses, this item is too similar to other items: it is more highly correlated than expected in terms of response pattern so in a sense it appears to provide some redundant information.

Given the small magnitude and the relatively small number of persons in the trial sample, it was decided to leave this item in the test until further evidence is obtained. In subsequent test development it may be appropriate to add more calculation items rather than infer calculation skill from a smaller number of items.

Item (K1) is to the right of the right-hand dotted line. It is implied that this item is measuring something different: it is less highly correlated in terms of response pattern than expected. It appears that this item is measuring something different from the other items. This item differed from the items nearby because it required *two* zones to be touched – the task required two piles of blocks that had the same volume to be identified (Figure 8). Given the small magnitude and the relatively small number of persons in the trial sample, it was decided to leave this item in the test until further evidence is obtained. At a further stage of test development more items requiring similar two-zone responses to be made could be added. Item U1 (Figure 9) proved to be very difficult although most persons could

make a rational response. It appears that the difficulty was due to the response set to touch regions, whereas the item would have been better as a graph completion task. The idea or concept being assessed is appropriate, however the response mode should change before the item is used.

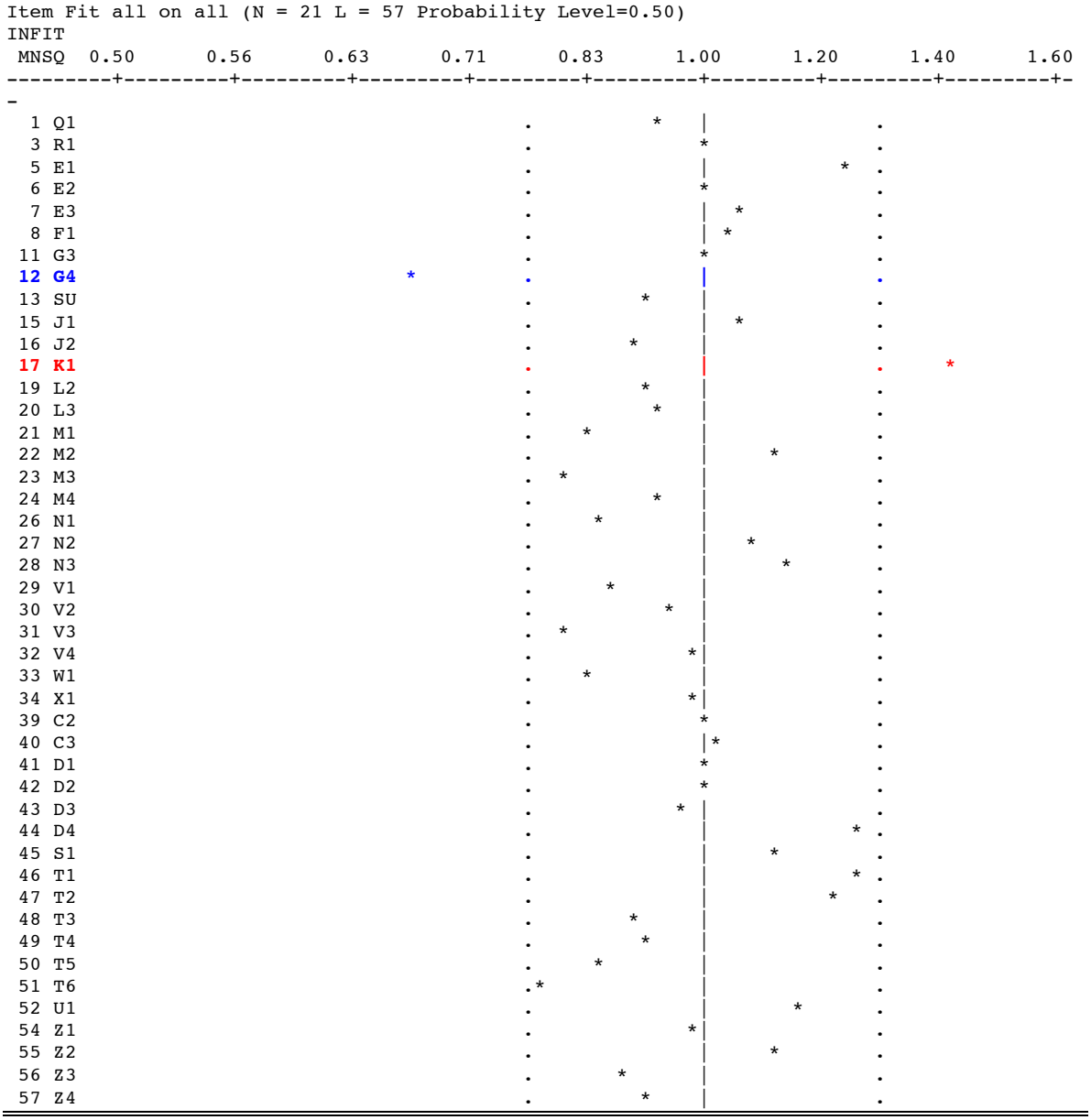


Figure 7 **Item Fit Map for Test with *Safety*, *Magnitude*, *Food*, *Identification*, and *Consumer Awareness* Tasks (Run 1)**

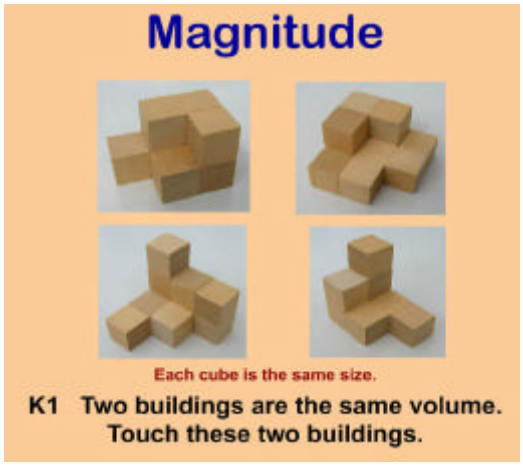


Figure 8 Item K1 requiring choice of 2-zones

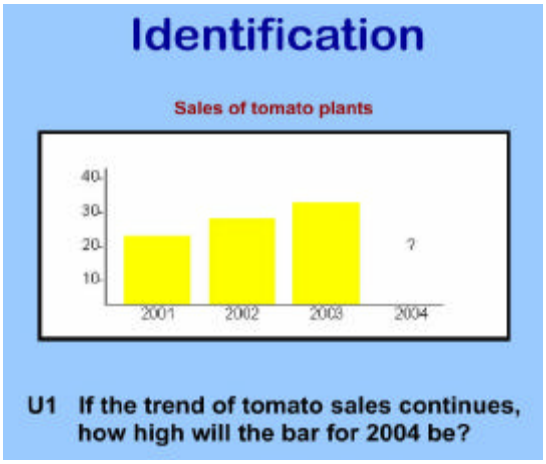


Figure 9 Item U1 needing response-mode alteration

Table 5 provides the numerical values of the thresholds for items marked 0 or 1 and for partial credit items. The former items have one threshold and the latter have more than one threshold (see the items on the right hand side of Figure 6). The proportion of students correct on each of the items marked 0 or 1 is shown in the second column of Table 5. The numeral in the item "Score" column is the number of students correct on that item. For example, 16 of the 21 students were correct (scored 1) on item Q1. For partial credit items the maximum score for the item is a multiple of 21. Where there are two thresholds such as item L3 the maximum score is 42 (= 2 x 21) and for three thresholds the maximum is 63 (= 3 x 21). For partial credit items, the numeral in the item "Score" column is the sum of student scores on that item. For item L3 (item 20 in Table 3) that sum of student scores for the item is 16.

Table 3 Item Estimates for Test with *Safety*, *Magnitude*, *Food*, *Identification*, and *Consumer Awareness* Tasks (Run 1)

Item Estimates (Thresholds) In input Order
all on all (N = 21 L = 57 Probability Level=0.50)

	ITEM NAME	SCORE	MAXSCR	THRESHOLD/S 1	2	3	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
1	Q1	16	21	-0.19 .55			0.93	0.90	-0.2	-0.1
2	Q2	0	0	Item has perfect score						
3	R1	20	21	-2.13 1.04			1.00	0.59	0.3	0.1
4	R2	0	0	Item has perfect score						
5	E1	16	21	-0.19 .55			1.24	1.30	0.9	0.7
6	E2	18	21	-0.88 .65			1.00	1.11	0.1	0.4
7	E3	19	21	-1.36 .77			1.06	0.80	0.3	0.1
8	F1	18	21	-0.88 .65			1.04	0.99	0.2	0.2
9	G1	0	0	Item has perfect score						
10	G2	0	0	Item has perfect score						
11	G3	15	21	0.09 .52			1.00	0.90	0.1	-0.1
12	G4	7	21	1.93 .50			0.67	0.57	-1.8	-1.3
13	SU	11	21	1.03 .48			0.91	0.93	-0.5	-0.1
14	H1	0	0	Item has perfect score						
15	J1	18	21	-0.88 .65			1.07	0.80	0.3	-0.1
16	J2	18	21	-0.88 .65			0.89	0.62	-0.1	-0.4
17	K1	4	21	2.77 .59			1.41	3.07	1.2	2.5
18	L1	0	0	Item has perfect score						
19	L2	18	21	-0.88 .65			0.91	0.62	-0.1	-0.4
20	L3	16	42	0.13 .97	4.19 1.49		0.93	0.92	-0.1	-0.1
21	M1	20	21	-2.13 1.04			0.84	0.30	0.1	-0.3
22	M2	19	21	-1.36 .77			1.12	1.54	0.4	0.8

Table 3 Item Estimates for Test with *Safety*, *Magnitude*, *Food*, *Identification*, and *Consumer Awareness* Tasks (Run 1) (continued)

Item Estimates (Thresholds) In input Order
all on all (N = 21 L = 57 Probability Level=0.50)

	ITEM NAME	SCORE	MAXSCR	THRESHOLD/S 1	2	3	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
23	M3	14	21	0.34			0.81	0.71	-0.9	-0.7
				.50						
24	M4	33	42	0.00	0.19		0.93	0.99	-0.1	0.3
				.97	.95					
25	M5	0	0	Item has perfect score						
26	N1	19	21	-1.36			0.85	0.48	-0.1	-0.4
				.77						
27	N2	35	42	-0.23	-0.04		1.09	7.47	0.3	3.1
				1.03	1.04					
28	N3	52	63	-0.89	-0.51	0.59	1.15	1.19	0.5	0.5
				1.27	1.20	.96				
29	V1	8	21	1.70			0.86	0.85	-0.7	-0.4
				.49						
30	V2	14	21	0.34			0.94	0.87	-0.2	-0.2
				.50						
31	V3	10	21	1.25			0.81	0.76	-1.1	-0.8
				.48						
32	V4	5	21	2.46			0.97	1.32	0.0	0.8
				.54						
33	W1	20	21	-2.13			0.84	0.30	0.1	-0.3
				1.04						
34	X1	11	21	1.03			0.99	0.93	0.0	-0.1
				.48						
35	A1	0	0	Item has perfect score						
36	B1	0	0	Item has perfect score						
37	B2	0	0	Item has perfect score						
38	C1	0	0	Item has perfect score						
39	C2	20	21	-2.13			1.00	0.59	0.3	0.1
				1.04						
40	C3	19	21	-1.36			1.03	0.70	0.2	-0.1
				.77						
41	D1	20	21	-2.13			1.00	0.59	0.3	0.1
				1.04						
42	D2	20	21	-2.13			1.00	0.59	0.3	0.1
				1.04						
43	D3	15	21	0.09			0.96	0.87	-0.1	-0.2
				.52						
44	D4	4	42	2.72	3.34		1.25	1.10	0.6	0.4
				1.28	1.41					

Table 3 Item Estimates for Test with *Safety*, *Magnitude*, *Food*, *Identification*, and *Consumer Awareness* Tasks (Run 1) (continued)

Item Estimates (Thresholds) In input Order
all on all (N = 21 L = 57 Probability Level=0.50)

	ITEM NAME	SCORE	MAXSCR	THRESHOLD/S 1	2	3	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
45	S1	17	21	-0.50			1.12	1.97	0.4	1.5
				.59						
46	T1	28	42	-1.06	1.42		1.27	1.23	1.0	0.7
				1.16	.90					
47	T2	19	21	-1.36			1.22	2.73	0.6	1.6
				.77						
48	T3	20	21	-2.13			0.89	0.37	0.2	-0.2
				1.04						
49	T4	17	21	-0.50			0.91	0.69	-0.2	-0.4
				.59						
50	T5	12	21	0.80			0.84	0.79	-0.9	-0.6
				.48						
51	T6	4	21	2.77			0.79	0.55	-0.6	-0.7
				.59						
52	U1	2	21	3.62			1.16	2.94	0.5	1.7
				.76						
53	Y1	0	0	Item has perfect score						
54	Z1	5	21	2.46			0.97	1.03	0.0	0.2
				.54						
55	Z2	5	21	2.46			1.12	1.50	0.5	1.1
				.54						
56	Z3	19	21	-1.36			0.88	0.48	0.0	-0.4
				.77						
57	Z4	19	21	-1.36			0.91	0.66	0.0	-0.1
				.77						
Mean				0.00			0.99	1.14	0.0	0.2
SD				1.66			0.15	1.15	0.5	0.8

Each threshold has an associated error. For example, item Q1 has a threshold of –0.19 with an associated error of 0.55. This error can be interpreted as suggesting a range for the threshold if the test is given again: –0.19 – 0.55 (= –0.74) to –0.19 + 0.55 (= +0.36). Similarly for partial credit items there are several ranges. For example, item L3 (see item 20) has thresholds of 0.13 with an associated error of 0.97 and 4.19 with an associated error of 1.49. These errors can be interpreted as suggesting two ranges for the thresholds if the test is given again: 0.13 – 0.97 (= -0.84) to 0.13 + 0.97 (= 1.10) and 4.19 – 1.49 (= 2.70) to 4.19 + 1.49 (= 5.68). The magnitude of these errors is a direct consequence of the small number of persons attempting the task (N = 21).

The magnitudes of the infit and outfit t-values shown in the last two columns of Table 3 are not extreme (in a statistical significance context). It was decided that all items should be retained until sufficient evidence was available to allow a more rigorous analysis.

Analysis of the test item data, using the Quest interactive test instrument (Adams and Khoo, 1996), showed a clear differentiation of test-item difficulty and user skill levels. In summary, 56 right/wrong and partial credit items (with a maximum possible score of 64), suitable for the evaluation of educational and academic skills of young adults, were able to be calibrated to form the basis of assessing progress made towards returning to study or vocational training. One item (U1) was identified as requiring an altered response mode.

Conclusion

This paper has described a research project for building an assessment system prototype. The project delivered an easy-to-use, effective skills-differentiation prototype system. In the longer term, this project paves the way for a more refined system, perhaps delivered as a fully interactive Web-enabled database. The instrument was not seen as threatening and generally the participants appeared to enjoy the testing experience. The trial of the system of presenting tasks on a Touch Screen was successful, and a sound basis for further development has been established. Further new items can be prepared and added to the existing pool of calibrated items provided that they are administered with existing items to suitable members of the client group. This project provided an innovative use of information communications technology services to assess the individual educational achievement levels of members of this at-risk learner group.

References

Adams, R.J.and Khoo, S.T. (1996). *QUEST*: *The interactive test analysis system*. (2nd Ed.). Melbourne, Victoria: Australian Council for Educational Research.

Gagne, R.M. (1985). *The conditions of learning: and the theory of instruction*. New York: Holt, Rinehart and Winston.

Izard, J.F. (2002). Describing student achievement in teacher-friendly ways: Implications for formative and summative assessment. In F. Ventura & G. Grima (Eds.) *Contemporary Issues in Educational assessment*. (pp. 241-252). MSIDA MSD 06, Malta: MATSEC Examinations Board, University of Malta for the Association of Commonwealth Examinations and Accreditation Bodies.

McKay, E. and Izard, J. (2004) Automated Educational Skills Evaluation: A Systems Design Case Study. Paper accepted for the 2004 International Conference on Computers in Education (30 November – 3 December, 2004): *Knowledge acquisition for life-long learning through human-computer interactions: creating new visions for the future of learning*. Melbourne, Australia (In press)

Merrill, M. D. (2003). Does Your Instruction Rate 5 Stars? (Keynote Address) In E. McKay, (Ed.) *eLearning Conference on Design and Development: Instructional Design - Applying first principles of instruction*. (pp. 13-14). Melbourne, Informit Library: Australasian Publications On-Line.

Okamoto, T., Matsui, T., Inoue, H. and Cristea, A. (2000). A Distance-Education Self-Learning Support System Based on a VOD Server. In C. Kinshuk, C. Jesshope and T. Okamoto, (Eds.) *International Workshop on Advanced Learning Technologies (IWALT 2000): Advanced Learning Technology: Design and Development Issues*. (pp. 71-72). Palmerston North, New Zealand: IEEE Computer Society.

Preece, J. (1994). *Human-Computer Interaction*. Harlow, England: Addison-Wesley.