Developing an innovative computer-based test

KASAMAPORN MANEEKHAO, NATJIREE JATURAPITAKKUL, RICHARD WATSON TODD and SAOWALUCK TEPSURIWONG – King Mongkut’s University of Technology Thonburi, Thailand

ABSTRACT
This paper examines the development of an innovative item type for a computer-based test. The test is one specifically designed for engineers and technicians in Thailand and includes a section assessing the reading of technical manuals. Having set test specifications, it was decided to test the reading of manuals in an authentic way by asking test-takers to assemble objects on the computer screen following given instructions. MacroMedia Flash MX was used to write this section. Based on feedback on this test, which was designed to test the ability to follow instructions, from both colleagues and students who had piloted the test, adaptations were made and test characteristics were investigated. The process of developing the innovative item type exhibits features from the fields of test design and materials design.

Introduction
Test design is perhaps the most complex and difficult of the ancillary responsibilities that teachers have to deal with outside the classroom. Because of this, in designing tests, many teachers use existing tests as a prime source of input for the new test design (Paltridge 1992; Alderson, Clapham and Wall 1995), especially as a source of inspiration for designing test items. Such input can be supported by checklists of do’s and don’t’s for the chosen item type (such as cloze, multiple-choice and essay item types) from published textbooks (for example, Forsyth, Jolliffe and Stevens 1999; Brown and Hudson 2002; Coombe and Hubley 2005). Where the item type being designed is innovative, however, existing tests can provide no input and textbooks can provide little guidance. Clearly, a different approach is necessary, and this article examines the procedures used to design an innovative item type for a computer-based test of English.

Test design
Several authors have suggested models of the test design process (for example, Carroll and Hall 1985; Alderson, Clapham and Wall 1995; Brown 2001).
Most of these models comprise four main stages in test design (although the terms used for the stages differ):

- planning – setting the purpose of the test and drawing up a table of specifications
- development – drafting and editing the test items
- piloting – trialling the test, final editing and establishment of validity and reliability
- operation and monitoring – administering the test and improving it for future versions.

Although the actual test design is likely to be an iterative non-linear process (Bachman and Palmer 1996; Weigle 2002), a model of the stages provides a framework to guide test designers. In the context of this article, it also allows us to focus on those aspects of the design of an innovative item type that are likely to provide the most useful insights.

In this paper, we will focus primarily on the development stage of test design for one specific section of a test. Before we can look at this, however, an understanding of the context and purpose of the test is needed. For this, we will look briefly at the planning stage first.

**Test of English for Thai Engineers and Technicians (TETET)**

In 2001, King Mongkut’s University of Technology Thonburi (KMUTT), a respected Thai government university, decided to fund a test of English proficiency designed specifically for engineers and technicians in Thailand. To run the TETET project, a team was set up with the goal of producing a valid and reliable test of English proficiency, which could be used as follows:

- by companies in recruitment, promotion and allocation of work responsibilities
- by universities in program evaluation and diagnosis of graduating students.

With these goals in mind, needs analyses through questionnaires and interviews were conducted with four groups of stakeholders, namely, practising engineers and technicians, final-year undergraduate students, senior university administrators, and senior administrators of the largest Thai engineering and technical companies. From these needs analyses, specifications of the test could be drawn up.

The key findings from the needs analysis process that are relevant to this article are as follows:
• all four language skills should be covered equally on the test
• the test should be computer-based
• one section of the test should focus on reading and understanding technical manuals.

The first two of these points suggest a certain amount of innovation in the overall test design. While speaking ability has previously been tested using computers (for example, the test run by the Community English Program at Columbia University), it is still not common. For TETET, spoken language was elicited from candidates by computer and recorded, but, given the limitations of computers, these recordings were human-marked.

While the overall test design contains an element of innovation, in this paper we will focus on the innovativeness of one of the sections of the test. This section concerned reading and understanding technical manuals, and the original table of specifications for this section is shown in Table 1.

**Table 1**: Table of specifications for the test section concerning reading technical manuals

<table>
<thead>
<tr>
<th>Reading: Technical manuals</th>
<th>Time</th>
<th>18 minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item type</td>
<td>Short-answer</td>
<td></td>
</tr>
<tr>
<td>Input</td>
<td>Page from manual</td>
<td></td>
</tr>
<tr>
<td>Reading strategy</td>
<td>Inference</td>
<td></td>
</tr>
<tr>
<td>Cognitive level</td>
<td>Synthesis</td>
<td></td>
</tr>
<tr>
<td>Number of questions</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Marking</td>
<td>Human-marked against model answers</td>
<td></td>
</tr>
</tbody>
</table>

**Initial section development**

Having set specifications for the section of the test, a selection of technical manuals was collected. These were analysed for length, language difficulty including level of technicality, complexity of the content and bias towards certain fields of engineering. Based on these criteria, many of the manuals proved unsuitable. The few remaining manuals, moreover, did not allow the writing of inferencing questions involving the cognitive level of synthesis.

Given that the reading of technical manuals was the key issue, the team decided to focus on the most appropriate manual and to examine how the key linguistic and discourse aspects of the content could be tested. Based on this, the table of specifications could then be revised (together with revisions to the specifications of other sections of the test to account for the changes).
Focusing on a manual concerning the installation of a water ioniser, it was decided to design a computer-based test whereby test-takers would have to manipulate objects on the screen following the instructions. In other words, we decided to test the ability to read technical manuals through a test designed to assess the ability to follow instructions (hereafter called ‘a test of following instructions’).

We believe that this test section has two key advantages. First, the fact that it can be computer-marked increases practicality. Second, since the primary purpose of reading technical manuals is to manipulate objects in the real world, such as operating machines or constructing appliances, a test which involves manipulating objects could be claimed to be authentic, even if a simulation. Thus it would have greater validity than tests that did not include manipulation. We also believe that the item type of manipulating objects is an innovative one. Although such an item type is used frequently both in testing the psychomotor skills of young children and in tests in some design subjects, we have not been able to find any records of previous work on tests involving the manipulation of objects to assess language skills. Given the novelty of the item type, it is worth examining the process of test design in more detail.

**Designing a test of following instructions**

The first stage in designing a test of following instructions was to create a paper-based version of the test section. This involved some rewriting and paraphrasing of the original manual and the design of new pictures. In addition, distracters were added and the process described in the manual was broken down into ten steps to allow the assignment of ten marks for the section.

From this paper-based version, a computer-based version was designed, initially using *Tool Book II Instructor Version 8.5*. However, computing problems, including the size of files, the need for the installation of a plug-in program and the need for all sections of the test to be compatible, meant that the authoring program needed to be changed to *MacroMedia Flash MX*. This program has been used previously to design computer-based tests (see Nezhad 2004), but these have been of the traditional multiple-choice and gap-filling item types. For TETET, however, we attempted to use the *Flash* authoring program for more innovative purposes.

A screenshot of the first attempt to design a computer-based test of following instructions is shown in Figure 1. In this example, test-takers can click on various parts of the screen and drag-and-drop objects to other parts of the screen.
This initial version of the test section was then critiqued by the TETET team members and revisions made based on this feedback. Two main revisions were made. The first concerned the addition of more objects and possible movements, since some items in the first draft contained too few possibilities and thus too few distracters. For example, the item shown in Figure 1 could be viewed as a polar item with a choice of dropping the aerator in either the toolbox or the dustbin. The second revision involved the addition of an explanation and labelled pictures at the beginning of the test section to avoid problems of test-takers misunderstanding the overall purposes and the names of the various components. With these revisions, the second version of the test section was ready.

**Developing a test of following instructions**

Following the design of a test, as we saw above, common practice is to pilot the test. While we next asked students to take and give feedback on the test, this step should be seen as still being part of the development stage of test design, rather than as part of the piloting stage. Although we used the data collected to analyse the level of difficulty of the test section, we were not attempting to establish test validity and reliability at this point. Rather, we were conducting a preliminary examination of how students as potential test-takers reacted to the innovative item type, and we were prepared to undertake major revisions (or even to start all over again) depending on the feedback received.

The second version of the test section was informally trialled with 51 technology students from KMUTT. They were given 15 minutes to com-
complete the test section. While they were working, two TETET team members observed their work and noted down anything interesting concerning how they completed the test. After they finished the test section, a questionnaire consisting of questions using a 5-point rating-scale (where 5 = satisfactory and 1 = unsatisfactory) was distributed, and the subjects were encouraged to give comments concerning their reactions to the test.

As we were interested in how the subjects would react to the item type of following instructions, the questionnaire covered two main areas. The first concerned the level of difficulty and clarity of the instructions, aspects traditionally investigated in test piloting. The second concerned the appearance and ease of use of the test, aspects not so frequently investigated in test design. The results are given in Table 2.

<table>
<thead>
<tr>
<th>Questionnaire statement</th>
<th>Average rating $\bar{X}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>The content level is suitable for people with a bachelor’s degree in science, engineering or technology</td>
<td>3.57</td>
</tr>
<tr>
<td>The language level is suitable for people with a bachelor’s degree in science, engineering or technology</td>
<td>3.63</td>
</tr>
<tr>
<td>The types of activities are suitable for the content</td>
<td>3.67</td>
</tr>
<tr>
<td>The instructions are clear, precise and easy to understand</td>
<td>3.76</td>
</tr>
<tr>
<td><strong>Average rating for the first section</strong></td>
<td><strong>3.66</strong></td>
</tr>
<tr>
<td>The program is attractive</td>
<td>3.86</td>
</tr>
<tr>
<td>The program is easy to read</td>
<td>4.14</td>
</tr>
<tr>
<td>Appropriate colours are used</td>
<td>4.14</td>
</tr>
<tr>
<td>The information in each page is adequate</td>
<td>3.92</td>
</tr>
<tr>
<td>It takes little time to understand how to operate the program</td>
<td>3.53</td>
</tr>
<tr>
<td>The program is easy to use</td>
<td>3.71</td>
</tr>
<tr>
<td><strong>Average rating for the second section</strong></td>
<td><strong>3.88</strong></td>
</tr>
</tbody>
</table>

The questionnaire results show that the subjects’ reactions were generally positive. There were, however, some discrepancies between the findings from the questionnaire and the findings from the observation and subjects’ comments. These concerned the ease of understanding the procedures for completing the test and the level of difficulty.

Although on the questionnaire the subjects rated the test easy to use and the instructions easy to understand, from the observation it was apparent
that most of the subjects were initially confused by the test. Many of the subjects attempted the test twice within the allotted 15 minutes. The general pattern was that on their first attempts they took around ten minutes to score 5 or 6 marks, but on their second attempts they could score 8 or 9 marks within three minutes. Their initial confusion was also evident from their comments:

At first, I did not understand how to use the program. When I read the instructions, I did not know if I had to click on an object or drag it.

At first, I did not know what to do. I did not know how to connect things.

There should be a pop-up telling if an object is movable when the cursor points at that object.

Once the subjects understood how to work on the program, they generally viewed the test as too easy (again in contrast to their questionnaire responses).

I think the test is too easy for people with a bachelor’s degree and the time should be reduced by two minutes.

The language is not suitable because it is too easy. Somebody just clicked without understanding and got marks.

There should be more choices because those who do not understand the instructions can possibly complete the test correctly.

The key finding from this trialling of the test section, therefore, suggested that major revisions were necessary. It was apparent that understanding how to complete the test was more important than language proficiency, and that luck could play an important role for test-takers.

One final comment from a subject suggested a direction that could be taken in revising the test. Focusing on the test authenticity, the subject suggested:

The test should be similar to the real situation in a workplace. For example, all objects and instructions are given at the same time, and a space is given for the test taker to put things together.

**Redeveloping a test of following instructions**

From the initial trialling of the test section, it was clear that extensive revisions were necessary. While the subjects’ positive reactions suggested that another attempt should be made along the same lines, how to work through the test of following instructions needed to be clarified so that the problem of understanding the procedures would not affect the test-takers’ scores, and
the number of distracters needed to be increased substantially to reduce the impact of guessing.

A further influence on our revisions was the need to prepare multiple versions of the test section in the future. Finding a range of suitable technical manuals that could be divided into ten separate sections was deemed impossible. It was therefore decided to use two or three separate manuals, each with fewer steps, to make up the ten marks. This decision also allowed us to reduce the impact of guessing, since less predictable steps could be chosen for inclusion on the test.

To facilitate understanding of the testing procedures and to increase authenticity, it was decided to completely reorganise the interface. Instead of the single frame shown in Figure 1, the screen would be divided into three areas as shown in Figure 2. One area contains the task input or manual instructions, a second area contains components to be added following the instructions (including some irrelevant components as distracters) and the third area is the work area where test-takers put the components together. With a ‘back’ button to allow test-takers to recheck the overall purpose and the names of the components, it was hoped that this set-up would enhance understanding of the test-taking procedures and increase authenticity.

![Figure 2: A screenshot of the third version of the computer-based test of following instructions](image)

**Piloting a test of following instructions**

At the time of writing, the test has been piloted in two ways. First, five subjects worked through the test while being observed. They were interviewed after
completing each stage of the test concerning their reactions to the item types, problems they had and suggestions for improving the test. Second, 20 further subjects took the test under test conditions to allow initial score analysis to be conducted.

The main purpose of the observed piloting with five subjects was to get feedback on the test. Thus, the data collected included suggestions for improving the test, such as the need to include a link allowing test-takers to re-read the instructions for the test of following instructions. More pertinently for the purposes of this paper, the subjects also gave more general comments on both the whole test and each section. They stated that they believed that the test was measuring the skills required by working engineers and technicians, suggesting that TETET has face validity. They also said that they found taking the test interesting and even enjoyable. This last comment particularly applied to the test of following instructions, which one subject stated had an almost game-like nature.

For purposes of test validation, the scores obtained from the piloting with 20 subjects allowed us to check the reliability of the human-marked sections of the test, compare the results from the different sections of the test and analyse the level of difficulty of each of the sections. For the test of following instructions, 10 marks were assigned, and subjects’ scores ranged from 0 to 10, with a mean of 5.90 and a standard deviation of 3.27. These basic findings suggest that the section is of roughly the right level of difficulty and distinguishes sufficiently between different subjects.

To check the reliability of the test of following instructions, we can compare the scores obtained on this section both with scores obtained on other reading sections and with overall test scores. We can also examine the section discrimination by comparing the scores on the section of the top ten subjects in terms of overall scores with the scores of the bottom ten subjects.

There is a problem in checking reliability with the other three reading sections of the test in that two of these other sections were also somewhat innovative. For example, one of the other sections involved reading from the Internet and included a large false website replete with internal links from which test-takers were required to find specific information. However, one reading section followed the traditional multiple-choice format for testing reading, allowing the test of following instructions to be compared with a traditional test of reading. This comparison showed that the scores on the two sections had a significant positive correlation ($r = 0.487; p < 0.05$). A similar significant positive correlation was also found when comparing the scores from the test of following instructions with the overall test scores ($r = 0.531; p < 0.01$).
To see if the test of following instructions discriminates between the subjects who scored higher overall and those who scored lower, the subjects were assigned into two groups based on their overall test scores, and the scores of these two groups on the section assessing the ability to follow instructions were compared using point biserial correlation. Again, we found a significant positive correlation ($r_{pbi} = 0.427; p < 0.05$), suggesting that the test of following instructions discriminates well between better and weaker test-takers.

**Discussion**

The test design described in this paper has features of interest in terms of both product and process. In terms of product, the item type of manipulating objects on the screen is practical for marking (although the amount of work needed to design the item type reduces overall practicality), is reliable since marking is objective, and has at least some element of validity as shown by the results of the initial piloting. As a general rule, the three test characteristics of validity, reliability and practicality stand in opposition to each other, so that adapting an item type to increase one of these characteristics often results in a reduction in the other two (Weir 1990; Johnson 2001). We are pleased to note therefore that the innovative item type we have chosen appears to promote all three of these characteristics, at least to some extent. For other characteristics of test usefulness (see Bachman and Palmer 1996), we believe that the item type has high authenticity and the test-takers’ reporting of enjoying taking the test implies that the item type has some positive impact. The innovative item type of computer-based manipulation of objects to assess the ability to follow instructions therefore appears to be a promising item type with high levels of overall test usefulness. In this way, we hope that we have shown that computer-based testing offers far more than just being a more convenient mode of doing what has previously been accomplished through paper-and-pencil-based tests. Rather, the power and flexibility of computers should be used creatively to add to the current limited list of item types derived from paper-and-pencil-based tests.

Regarding the test design process, we have to some extent followed the standard practices employed in models of test design. At a global level, we have worked through the stages of planning and development and are at present conducting piloting. However, because the item type chosen to test the reading of technical manuals is innovative, we have taken a much longer time and engaged in more processes in the development stage than is probably the norm in most test design. Even in this longer than normal development stage, it could be argued that we have simply been doing what
other test designers recommend. For example, getting committees to review first drafts of a test (Alderson, Clapham and Wall 1995) and getting feedback on test design from potential test-takers (Chamberlain 1996) have been previously incorporated into the test design process.

However, the processes undertaken in the test development stage also have clear parallels with another field that is not normally directly linked with test design, namely, materials design, adaptation and evaluation. As with test design, several models have been suggested to guide materials design, adaptation and evaluation. Two are of particular relevance to the TETET project.

First, Allwright (1990) argues that implementing materials in teaching follows a cyclical process as follows:

- decision
- organisation
- implementation
- monitoring
- evaluation
- decision, and so on.

In developing a test section, we first made a decision to test the reading of technical manuals through following instructions. We then organised a test and implemented it in trialling, which was monitored and evaluated. The findings from this evaluation allowed us to come to a new decision concerning the test section.

Similarly, Watson Todd (1999) also proposes a cyclical model of materials modification starting from the identification of objectives and the initial production of materials. Following this is a cycle of trialling the materials, evaluation, identification of strengths and weaknesses, identification of principles to overcome weaknesses, and modification of materials. Again, the development of the test of following instructions could be viewed as following this cycle.

In these ways, the processes we have followed in developing an innovative item type could be based within the principles of materials design as much as within the field of test design. It seems likely that the applicability of models of the process of materials design to the present study is due to the innovativeness of the item type. Where test design involves innovative item types, therefore, test designers may find it useful to look beyond the testing literature for guidance.
Conclusion
In this paper we hope to have achieved two purposes. First, we hope that we have explained and illustrated an innovative item type – a computer-based test of following instructions – which other teachers may be able to use in their own tests. Second, we have investigated the processes undertaken in developing this item type. Given that the item type is innovative, as might be expected we have spent more time on the development stage of the test than is perhaps the norm. Furthermore, the processes followed in developing the test section can be viewed as involving materials development in addition to the standard practices of test design.

REFERENCES


