Web-based Assessment and Test Analyses (WATA) system: development and evaluation


*Jen-Teh Junior College of Medicine, Nursing and Management, Miaoli, Taiwan,
†Graduate Institute of Science Education, National Changhua University of Education, Changhua, Taiwan,
‡Department of Biology, National Changhua University of Education, Changhua, Taiwan and
§Department of Information Systems and Computing, Brunel University, Middlesex, UK

Abstract

This study introduces the development of a Web-based assessment system, the Web-based Assessment and Test Analyses (WATA) system, and examines its impacts on teacher education. The WATA system is a follow-on system, which applies the Triple-A Model (assembling, administering, and appraising). Its functions include (1) an engine for teachers to administer and manage testing, (2) an engine for students to apply tests, and (3) an engine for generating test results and analyses for teachers. Two studies were undertaken to assess the usefulness and potential benefits of the WATA system for teacher education. In the first study, 47 in-service teachers were asked to assess the functions of the WATA system. The results indicated that they were satisfied with the Triple-A Model of the WATA system. In the second study, 30 pre-service teachers were required to use the WATA system during the teacher-training program. After 4 months of experience in using the WATA system, the pre-service teachers’ perspectives of assessment have been changed significantly. The findings of these two studies might provide some guidance to help those who are interested in the development of Web-based assessment and intend to infuse information technology into teacher education.

Keywords

assessment, computer-based testing (CBT), WATA (Web-based Assessment and Test Analyses) system, web-based testing (WBT), pre-service teacher education, testing system

Introduction

The World Wide Web (Web) has been gaining both sophisticated users and general popularities within educational settings. It has become an inexpensive, easily accessible way to communicate, distribute information, teach courses, and conduct research (Lan 2001). However, the International Society of Technology in Education (ISTE 1999) released findings from its survey on information technology (IT) in teacher education, concluding that most student teachers do not use technology routinely during field experience and do not work under the guidance of master teachers and supervisors who can advise them on technology use. The survey also suggests that the student teachers’ classroom skills and the actual use of IT during teacher training are critical predictors for IT integration in the future. Hence, using IT in the teacher-training program has become an important issue.

Magnusson et al. (1999) suggested that teacher-training programs should consider including pedagogical knowledge (PK), content knowledge (CK), and pedagogical content knowledge (PCK), which suggests...
that the only e-mail services and Web pages for courses are not sufficient for the issue of integrating IT with teacher education. The teacher-training program really requires more Web tools for pre-service teachers to interact with the Web. Ahern et al. (2000) suggested that a Web-based assessment was a useful tool for teacher education. However, it is still not clear what kinds of structure a Web-based test system should be equipped with and what benefits it can have for teacher education.

In this paper, we describe how we develop a Web-based assessment and test analyses system, the Web-based Assessment and Test Analyses (WATA) system. Furthermore, two studies were conducted to examine the potential benefits of the WATA system on the teacher-training program.

Review of literature

Computer-based testing (CBT)

CBT (Huff & Sireci 2001) allows students to read test items on the computer screen, select answers with the mouse or keyboard, re-examine and revise them, and then send them out, and then log out when they are finished. In other words, the computer simply acts as the medium for students to take exams, for teachers to construct tests, and for the transmission of test papers. As for CBT, since the late 1970s, science educators have been experimenting with the use of microcomputers for the conceptual and attitudinal assessment of their students (Bork 1981; Arons 1984, 1986; Waugh 1985; Cole et al. 2001). Moreover, the multiple-choice, machine-scored, and standardised instruments have been developed to assess the conceptual knowledge of physics in the late 1980s (Hestenes et al. 1992; Cole et al. 2001). Recently, because Internet and database technology have been fully developed, CBT, which before was once hosted only on personal computers (PCs) or local area networks (LANs), has now gradually been upgraded to work on the Internet using browsers as the test interface so that users can use it anywhere in the world. Compared with CBT on a PC or LAN, designed and developed separately by IBM (Oblinger 1992) and Arthur Anderson Company (Dennis & Gruner 1992), Web-based testing (WBT) is more flexible and more useful to pedagogic research (Bonham et al. 2000). The name WBT was coined to distinguish it from the more conventional CBT, which only works on a single PC or LAN.

Varieties of CBT

CBT can be classified into two main classes, regular and adaptive, according to the timing of use and the differences in the underlying theoretical foundation (Mason et al. 2001). Regular CBT (R-CBT) is simply a paper-and-pencil test, which is carried out on a computer. The assessment items for students are identical (Mason et al. 2001). Conversely, the test items in adaptive CBT (A-CBT) vary with testees’ answers. Test papers thus reflect the abilities of testees. A-CBT was developed on the basis of the Item-Response Theory (IRT), which emphasises that ‘the probability that a person answers an item correctly depends only on the person’s ability’ (Hambleton & Swaminathan 1985). The basic assumption of IRT is that ‘a person of higher ability is more likely to give the correct response’ (Green 1991, p 246). In this way, persons of varying ability (as measured by their answers to previous questions) are given questions that are most likely to accurately assess their knowledge of a given matter (Wise & Plake 1990). The main benefits of A-CBT are testing efficiency and accuracy in identifying cognitive ability or ability limits (Mason et al. 2001).

In comparison with R-CBT, A-CBT is more useful for modern education practices because it places greater emphasis on individual differences and materials-based teaching by providing an adaptive computer-based assessment environment (Mason et al. 2001). Although A-CBT can play the role of a diagnostic assessment tool, many researchers still resist application of A-CBT in a variety of assessments (Bugbee 1996). One of the problems is that A-CBT relies on IRT. IRT in turn requires that each test item be administered to a large group of respondents in order to determine the response function for each item. This is then used to determine the ability score of each respondent. In this way, A-CBT requires a large population of students, so it is not currently feasible for small-scale classroom tests (Mason et al. 2001).

Web-based testing (WBT) system

With the widespread application of computer technology to instruction in recent years, the use of computers to administer tests is increasingly applied in education (Bonham et al. 2000; Buchanan 2000;
In addition, because of the rapid advancement of Internet technology, many assessment systems are designed based on WBT, which works over the Internet and utilizes browsers as its interface. WBT systems not only provide online real-time tests and online construction of items, but also enable teachers to check answer sheets rapidly and record scores over the Internet (Bonham et al. 2000). Moreover, since these systems operate in Web-based environments, audio, video, virtual, and dynamic pictures designed by JAVA scripts or FLASH may be incorporated into tests.

A number of studies suggest some required features and functions for WBT or CBT. Alessi and Trollip (1991) stated that the key features of CBT should be ‘establishment of item pools, construction and administration of tests or exams’. Furthermore, Gardner et al. (2002) developed a computer-supported learning system, named CECIL, which included an interesting function: ‘Self-Assessment’, to enhance students’ learning effectiveness. The function of ‘Self-Assessment’ is equipped with item pools and teachers can administer and construct examinations easily through the Internet. They also pointed out that the advantages of item pools are that ‘teachers are able to incorporate large item banks (item pools) from textbook publishers and batch load these questions with a minimum of manual effort’. Moreover, Gardner et al. (2002) also stated that teachers who administer and construct an exam through the Internet have the advantage of helping students to check their understanding of the learning materials at all hours.

Bonham et al. (2000) stated that it is important that WBT can be assessed through the Internet with standard browsers as students can access the online test anywhere and anytime. Bonham et al. (2000) and Randolph et al. (2002) discussed the security and equity issues of online tests. They both indicated that the password function for identification of testees is essential. Furthermore, Bonham et al. (2000) also stated that automatic grading and collection and recording of student scores are important for teachers because they can make teachers trace and manage students’ learning status more easily than traditional paper-and-pencil tests. In CECIL, the ‘Gradebook’ function can help teachers to record and update any student scores data at any time and to monitor students’ learning status. As for ‘collection and recording student scores’, Kumar et al. (1994) stated that because of the excellent capabilities of data management, collecting the processing of test data becomes much easier than ever; likewise it becomes possible to provide active feedback to the learning process and active detection of errors and misunderstandings. In addition to the functions above, the random function is also very important. Randolph et al. (2002) indicated that the use of random questions could keep students’ curiosity on the examination and then encourage them to try the quiz repeatedly, and thus drilling the students in key concepts.

**Development of the WATA system**

Wang et al. (2002) suggested that an assessment system should apply the Triple-A Model (assembling, administering, and appraising) as the baseline qualification in order to provide the most comprehensive form of CBT or WBT and to be more suitable for teacher education. The Triple-A Model was developed by Wang et al. (2002) based on the ‘basic steps in classroom testing’ proposed by Gronlund & Linn (1990, pp. 109–141, 228) and interviews with 17 in-service teachers and assessment experts. The Triple-A Model includes the essential functions of CBT or WBT discussed above (Alessi et al. 1991; Kumar et al. 1994; Bonham et al. 2000; Mason et al. 2001; Gardner et al. 2002; Randolph et al. 2002):

- **Assembling**: to construct item pools, test items, and schedules of tests.
- **Administering**: to assign the test items and item choices randomly to testees; provide personal identification numbers (PINs) and examination passwords for testees to apply the test through Web; to collect and record the scores data and all the process data of the tests.
- **Appraising**: to analyse the collected process data of tests and to generate the statistic report.

In this research, the WATA system is structured in accordance with the Triple-A Model. A more detailed explanation follows in the next section.

**WATA system software and operating requirements**

*Developing software for the WATA system.* The WATA platform was the FreeBSD running on an
Apache WWW Server. The core of the WATA system is programmed in PHP and Perl and works with MySQL to handle extremely large data sets and analytical programs efficiently.

Limitations on users. The WATA system is classified as WBT. Users have to use a browser to access the WATA system over the Internet.

Personal interface and Triple-A Model of the WATA system

Adhering to the concept of interface personalisation, the WATA system allows every teacher who uses it to have a personalised interface equipped with the Triple-A Model (see Fig. 1).

After teachers log in with their passwords, they can open their personalised Triple-A Model interfaces to the WATA system, and with the help of these unique interfaces they can arrange analyses and assessments. In Fig. 1, there are a series of Web tools concerning the Triple-A Model for teachers in the left column, such as ‘constructing tests with two-way charts’, ‘item feedback’, ‘item query’, ‘multi-exam schedule manager’, ‘test & item analyses’, ‘misconception analyses’, etc. The Triple-A Model of the WATA system contains the following functions.

Assembling

The assembling functions include ‘expert item pool’, ‘property serial number’, ‘constructing test with two-way chart’, ‘set-up’, and ‘arrangement of multi-exam schedule’.

‘Expert item pool’ and ‘property serial number’. The ‘expert item pool’ is divided into two parts: the ‘item feedback pool’ and the ‘system item pool’. The ‘item feedback pool’ provides teachers with zones to store all items. These items are designed or revised by the teachers, but cannot be utilised to construct real tests; i.e., they are candidate items for subject-matter experts to review. The ‘system item pool’ is used to store formal items for real tests. The items will be automatically assigned to ‘discriminating power (DP)’ index and ‘item difficulty (ID)’ index after students take the test. The relationship between ‘item feedback pool’ and ‘system item pool’ is shown below.

Complementing the ‘expert item pool’ is the concept of ‘property serial number’, under which each item is provided with a ‘serial number’ representing its properties. Such properties include ‘discriminating power’, ‘item difficulty’, ‘Bloom’s taxonomy of cognitive domain’, and ‘concept serial number’. The ‘property number’ system describes all properties of a given item in detail.

Fig. 1 Personalised interface of the WATA system.
The WATA system presents items with pictures, video, audio, animation, and other multimedia effects, gives simple feedback and explanation for each item, and offers real-time feedback when tests are completed. It is a practical and useful tool for construction of test items or data collection for education researchers.

Constructing tests with two-way charts (also known as the table of specifications). The WATA system adheres to the requirements of Guidelines for Computer-Based Tests on Performance and Interpretations published by the American Psychological Association (APA). The APA guidelines direct that the interpretation of scores must account for or remove any effects due to computer administration. Furthermore, the validity of a computerised test must be proven and publicised by developers. In order to determine the validity, defined by the equivalence of a computer-based test to a valid paper-based test, two criteria need to be met according to the APA guidelines:

... the rank order of scores of individuals tested in alternate modes closely approximate each other, and the means, dispersions, and shapes of the score distributions are approximately the same, or have been made approximately the same by rescaling the scores from the computer mode... (APA 1986, p. 18)

For traditional paper-and-pencil tests, the validity of the assessment is mainly defined as content validity. The most concrete method of determining the content validity is to use a two-way chart to decide details of proposed items and their quantity. The WATA system provides teachers with a two-way chart to edit the online test according to pre-determined objectives and learning concepts. The ‘concept serial number’, ‘Bloom’s taxonomy of cognitive domain’, and ‘item difficulty’ of each item, as well as the functions mentioned above, allow teachers to find all items corresponding most closely to the procedures of item construction in the test theory and to the demands made by APA and CBT (see Fig. 2).

Set-up and arrangement of multi-exam schedules. The personalised interface offers every teacher a unique WATA interface with its own Triple-A Model. Every teacher can freely construct items from the system item pool with the two-way chart and set up multi-exam schedules, which may be adjusted as necessary. In other words, teachers can administer many examinations in the WATA system and, in the meantime, can cancel, terminate, or temporarily halt exams. Additionally, methods for conducting Web-based assessments can be set for each examination, including time limits, examination passwords, testee PIN (which can be generated manually or automatically by the WATA system), test re-construction with the two-way chart, paper-and-pencil test, and duplicate test, as well as options that permit the

---

![Fig. 2 Construct test with two-way chart. (A) Concept serial number. (B) Bloom's taxonomy of cognitive domain. From left to right, they are 'knowledge', 'comprehension', 'application', 'analyses', 'synthesis', and 'evaluation'. (C) A button used to select proper items, which are sought out according to the specific 'concept serial number' and 'Bloom's taxonomy of cognitive domain' in the system item bank, to assemble a test. (D) The number of items sought out and selected. (E) A button used to delete disfavoured items. (F) Start out for generating the test and set the administering method.](image-url)
display of correct answers, automatic grade reporting by e-mail, permitting testees to check their own and others’ grades, display of instructor e-mail address for testee questions after the tests, and so on (see Fig. 3). There are three rows in Fig. 3, showing the exam names ‘eljh203’, ‘eljh204’, and ‘eljh205’, and four columns (A–D): A, ‘exam name’; B, ‘set’; C, ‘range’; and D, ‘exam date’. Column A includes the function of showing the schedule name, trying and deleting the schedule. Column B includes the function of setting the administering method, re-constructing the test according to the existing two-way chart, listing test contents, listing the test in the format of the paper-pencil test, deleting a specific testee’s answering information, modifying answers, and rescoring and duplicating the exam schedule. Column C shows chapters that the test content covered. Column D includes the functions of setting the schedule starting and stopping time, exam passwords and testee accounts.

**Administering**

The WATA system is equipped with a normal assessment system for students to take an examination, but the structure of the WATA system is tighter (see Fig. 4).

For administering tests, security is the most critical issue in a Web-based environment, since the Internet is an open environment. For preservation of the convenience and security of the Internet, the WATA system has four security features. First, testees have to key in the accounts of teachers who administer the test. Second, all testees must have special examination passwords provided by their instructors. Third, all testees have to key in their own PIN provided by the teacher before logging in and taking the test. Finally, all test items and item choices are randomly assigned, meaning that every testee will see a different test.

As for correcting tests, the WATA system can not only make records of grades just as other assessment systems, but also dump the process from testee answers for further analyses. Additionally, as students input their answers for corrections, the WATA system will inform students of their grades and performance on each item, and then send grade reports to the students and teachers by e-mail. If the system item pool could obtain simple feedback and explanation for each item, the WATA system would list these as real-time feedback for students to consult and enhance their learning. Testees can discuss questions with teachers through e-mail exchanges.

**Appraising**

The WATA system offers the option of ‘construct tests with the two-way chart’ to confirm the validity of the assessments, as well as allows the instructors to analyse data from test answers. The data will help

<table>
<thead>
<tr>
<th>Exam Name</th>
<th>Set</th>
<th>Range</th>
<th>Exam Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>eljh203</td>
<td></td>
<td>10--10</td>
<td></td>
</tr>
<tr>
<td>eljh204</td>
<td></td>
<td>10--10</td>
<td></td>
</tr>
<tr>
<td>eljh205</td>
<td></td>
<td>10--10</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 3 Set-up and arrangement of multi-exam schedules.
teachers appraise each piece of information. The foregoing functions ensure an assessment that will be applicable in future class instruction. The ‘test analyses’ in the WATA system is classified into ‘item analyses’, ‘test analyses’, and ‘misconception analyses’.

The WATA system analyses students’ grades and data from their answers to form ‘misconception analyses’, ‘test analyses’, and ‘item analyses’, subsuming listed information regarding ‘item difficulty’, ‘KR20 Reliability’, student score conditions, misconceptions, and so on. Such comprehensive analyses will be helpful for teachers to investigate the processes of student learning and their own instruction, and to understand whether their assessment is appropriate. The WATA system can appraise and analyse all students’ grades and answers in a short time after tests are finished. A report of the test analyses will be generated as HTML format and the content will include standard statistical descriptors such as ‘variance’, ‘standard deviation’, ‘testee T-score, z-score and Z-score’, ‘mean of all grades’, ‘average difficulty of the test’, ‘KR20 reliability’, ‘discriminating power (DP)’, ‘item difficulty (ID)’, ‘answers on each item in upper group and lower group’, ‘wrong answers of students for each item’, ‘distracter analyses of all testees, upper group and lower group’, and so on (see Fig. 5).

Fig. 4 Administering procedure of the WATA system.

Fig. 5 Summary table of item analyses and distracter analyses of all items. A–E: decisions made on each item; N/A: records testees who gave up answering; Num: how many people answered this item; DP: discriminating power; ID: item difficulty; and H: the percentage of testees in the upper group who checked the decision; L: testees in the lower group.
Additionally, the WATA system will automatically create a pie chart or a bar chart for teachers to understand testees’ scores distribution. As to the function of ‘misconception analyses’, the WATA system will provide percentages of error response on a specific concept according to the ‘concept serial number’ of the teacher’s ‘two-way chart’.

Comparing the WATA system with other WBT systems

Using Gronlund and Linn’s (1990) ‘basic steps in classroom testing’ as a framework for the comparison between the WATA system and several famous WBT systems, such as ‘LON-CAPA’ by Michigan State University (http://www.lon-capa.org/), ‘Question Mark’ (http://www.questionmark.com/tryitout.htm) by Question Mark Computing, ‘TopClass’ (http://www.wbtsystems.com) by WBT Systems (Bonham et al. 2000), Table 1 lays out the differences.

‘Step 1–Step 7’ in Table 1 are the ‘basic steps in classroom testing’ proposed by Gronlund and Linn (1990). The content of the ‘basic steps in classroom testing’ is as follows:

Step 1: Determining the purpose of testing.
Step 2: Constructing the two-way chart.
Step 3: Selecting appropriate items according to the two-way chart.
Step 4: Preparing relevant items.
Step 5: Assembling the test.
Step 6: Administering the test.
Step 7: Appraising the test, such as ‘variance’, ‘standard deviation’, ‘testee T-score, z-score and Z-score’, ‘mean of all grades’, ‘average difficulty of the test’, ‘KR20 reliability’, ‘DP’, ‘ID’, ‘answers on each item in upper group and lower group’, ‘wrong answers of students for each item’, ‘distracter analyses of all testees, upper group and lower group’, etc.

As shown in Table 1, most WBT systems lack ‘construct the two-way chart’ and ‘select appropriate items according to the two-way chart’ except for the WATA system. In addition, the ‘appraising the test’ function of the WATA system offers more functions than other WBT systems. For example, the LON-CAPA, TopClass, and Question Mark are merely present testees’ original scores, static tables of scores, static charts of scores and answering status of a specific item, without detailed test analyses and item analyses. The foregoing showed that the WATA system was relatively suitable for teacher education.

Evaluation of the WATA system

The major purpose of this study was to examine the potential benefits of the WATA system on teacher education; two studies were conducted to answer this question. These included a survey of in-service teachers’ satisfaction with the WATA system and an investigation on pre-service teachers’ perspectives about assessment.

Study 1

The first study was conducted to assess the user satisfaction of the WATA system. The User Satisfaction Survey of the WATA system (USSW) was used to survey in-service teachers’ satisfaction of the design of the Triple-A model in the WATA system.

Participants

Participants in this study consisted of 47 science teachers from 39 different schools. These teachers had been using computers and the Internet for at least 4 years, and all of them had their own websites for Web.
based instruction (WBI). They volunteered to participate in this study and understood the content and goals of this study. Participants included four university professors and 43 high-school teachers. Of the participants, four had doctoral degrees, four had master’s degrees, and 39 bachelor’s degrees. Of the participants, 57% were male and 43% of them were female. The average age of the teachers was 35.28 years (SD = 6.71). The average teaching experience of the university-professor group was 13.50 years (SD = 14.57), and the average teaching experience of high-school teacher group was 5.21 years (SD = 4.20).

Instrument

The USSW was used to evaluate users’ satisfaction with the detailed design of each ‘A’ (assembling, administering, appraising) in Triple-A of the WATA system. Two assessment experts and two IT experts reviewed the content of USSW. There were three sections in USSW: ‘assembling subscale’, ‘administering subscale’, and ‘appraising subscale’. All subscales used a five-point Likert scale, including ‘strongly satisfied (5 points)’, ‘satisfied (4 points)’, ‘neutral (3 points)’, ‘unsatisfied (2 points)’, and ‘strongly unsatisfied (1 point)’.

Assembling subscale’ had 27 items and the Cronbach α was 0.93. It was applied to evaluate the level of satisfaction with the design of ‘expert item pool’, ‘property serial number’, ‘constructing tests with the two-way chart’, and ‘set-up and arrangement of multi-exam schedule’. ‘Administering subscale’ had 21 items and the Cronbach α was 0.95. It was used to determine the level of satisfaction with the design of ‘safety and fairness’ and ‘real-time grading, feedback and asynchronous interaction’. ‘Appraising subscale’ had 20 items and the Cronbach α was 0.92. It was used to evaluate the level of satisfaction with the design of ‘test analyses’, ‘item analyses’, and ‘misconception analyses’.

Procedures

The research process consisted of three steps: 1 The WATA system was designed and programmed based on the test theory. This was completed by 31 July 2001. 2 Forty-seven participants used the WATA system for at least 1 month to familiarise themselves with the system and offered practical advice for improvement of the system. 3 The USSW was used to collect the quantitative data. Each participant was asked to answer each item in the USSW. All data were collected anonymously. The survey went on for about 7 months.

Results

There were 47 valid subjects. An average above 3.0 points represents a tendency towards satisfaction with the design. spss\textsuperscript{TM} Ver. 8.01 (SPSS Inc, Chicago) was used to analyse quantitative data. Table 2 shows the average scores given by 47 participants to the description of each item in ‘assembling subscale,’ ‘administering subscale’, and ‘appraising subscale’.

In the ‘assembling subscale’, participants gave high scores to the four major designs of ‘assembling’,

<table>
<thead>
<tr>
<th>Functions of WATA</th>
<th>Mean</th>
<th>SD</th>
<th>Cronbach α</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Assembling</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expert item pool</td>
<td>4.74</td>
<td>0.12</td>
<td>0.93</td>
</tr>
<tr>
<td>Property serial number</td>
<td>4.18</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>Constructing test with two-way chart</td>
<td>4.64</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>Set-up and arrangement of multi-exam schedule</td>
<td>4.77</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td><strong>Administering</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Safety and fairness</td>
<td>4.69</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>Real-time grading, feedback and asynchronous interaction</td>
<td>4.48</td>
<td>0.14</td>
<td>0.95</td>
</tr>
<tr>
<td><strong>Appraising</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test analyses</td>
<td>4.74</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>Item analyses</td>
<td>4.39</td>
<td>0.14</td>
<td>0.92</td>
</tr>
<tr>
<td>Misconception analyses</td>
<td>4.59</td>
<td>0.06</td>
<td></td>
</tr>
</tbody>
</table>
especially the ‘set-up and arrangement of multi-exam schedule’.
In the ‘administering subscale’, participants appeared to be satisfied with the two functions of ‘administering’. In addition, in the ‘appraising subscale’, it shows that participants were apparently satisfied with the three functions of ‘appraising’, especially the ‘test analyses’. In summary, users showed a high level of satisfaction with the functions of the WATA system.

Study 2

This study investigated the benefits of the WATA system as a tool for teacher education. The WATA system was used as a Web tool to construct, administer, and analyse a testing according to the current contents of the biology textbook of a high school in Taiwan. Furthermore, the Survey of Assessment Perspectives (SAP) was used to identify the changes of the pre-service teachers’ perspectives about assessment.

Participants

Participants were 30 pre-service biology teachers (16 males and 14 females) who had taken the course of ‘Measurement and Assessment in Teaching’ before. The average age of the pre-service teachers was 21.38 years (SD = 0.71). All participants possessed a reasonable degree of computer literacy, including familiarity with common word-processing programs, and all were able to access the Internet.

Instrument

SAP was used to assess the pre-service teachers’ perspectives of the assessment functions and procedures. It was designed by following the ‘basic steps in classroom testing’ (Gronlund & Linn 1990) and was reviewed by three experts in the areas of biology education and educational assessment. SAP is a nominal-categorical measurement and each item grouped participants into categories based upon agreement (Yes or No). ‘Yes’ is assigned the number of ‘1’ and ‘No’ is assigned the number of ‘0’. There are two subscales in SAP, named ‘perspectives about assessment functions’ and ‘perspectives about assessment steps’. The ‘perspectives about assessment functions’ had seven items and the Cronbach α was 0.71. The ‘perspectives about assessment steps’ had eight items and the Cronbach α was 0.78.

Procedure

This study was implemented in the course of ‘Biology Teaching Materials and Methods’, which is an important required course for pre-service biology teachers. The course was taught using a WBI. This WBI course was particularly developed by pre-service teachers, instead of professors. The study consisted of four steps:

1. At the beginning of the semester, the professor assigned 30 topics, which covered current contents of a high-school biology textbook to 30 pre-service teachers. Each participant was requested to create a WBI course according to his/her responsible topic, and to construct their own WATA item pools consisting of at least 50 topic-related items. In this step, the SAP survey was used to collect pre-service teachers’ preconceptions about assessment.

2. In the following week, each pre-service teacher who acted as a lecturer then had to use his own WBI materials to give a two-week lesson to 29 peers (the other 29 pre-service teachers) in turns. At the same time, each pre-service teacher also had to assemble and administer WATA formative assessments to peers in their WBI. The other 29 pre-service teachers had to enter the lecturer’s WBI website in their off-hours and they needed to conduct the WATA formative assessments repeatedly. The WATA system randomly selected ten items from the lecturer’s pre-constructed item pool, and assembled a formative assessment automatically. The other 29 pre-service teachers could repeatedly take the formative assessment in the WBI. In other words, the pre-service teachers who were responsible for the special topic acted as a lecturer for the other 29 pre-service teachers, who in turn played the role of learners.

3. After a 2-week lecture, each pre-service teacher, who was responsible for the special topic, had to use the WATA system to assemble a summative assessment to evaluate the learning effectiveness of their peers. In addition, each pre-service teacher also had to use the WATA system to perform test analyses, and item analyses and then had to compose an analyses report as homework to discuss the learning effectiveness of peers and how to improve his/her teaching and item-construction skill.

4. After 4 months, all pre-service teachers finished their WBI, and SAP was used to collect the post-test data.
Result
SAP is a nominal-categorical measurement and each item grouped participants into categories based upon agreement (Yes or No). ‘Yes’ is assigned the number of ‘1’ and ‘No’ is assigned the number of ‘0’. There are 30 valid samples. The SPSS™ Ver. 8.01 was used to perform statistic analyses. The data from SAP are transformed as percentage of agreement, and then Cochran’s Q with Monte Carlo simulation is used to test the significance of the pre-test and post-test data. The result of the subscale ‘perspectives about assessment functions’ (see Table 3) indicated that the 4 months training using the WATA system significantly transformed participants’ perspectives about assessment functions. After 4 months, participants were able to identify additional functions of the classroom assessment, especially ‘to analyse individual student misconceptions’, ‘to analyse and improve instructional strategy’, ‘to understand the strengths and weaknesses of test items’, ‘to understand the strengths and weaknesses of choices for a multiple-choice item’, and ‘to understand the strengths and weaknesses of a test’.

Moreover, the result of the subscale ‘perspectives about assessment steps’ indicated that the 4-month training using the WATA system significantly transformed the participants’ perspectives on assessment procedures. Participants used the WATA system to assemble, administer, and appraise classroom assessment in their WBI, which made them more familiar with the procedures of the classroom assessment and more willing to follow the standard procedures for administering a test. Table 4 shows significant transformations in all steps except for the three steps, required for every test: ‘assembling the test’, ‘administering the test’, and ‘scoring the test’.

Concluding remarks
In this research, the WATA system was constructed based on the Triple-A Model and was employed as a Web tool for teacher education. The Triple-A Model was concluded from the interviews of in-service teachers and assessment experts in Taiwan, as well as the ‘basic steps in classroom testing’ stated by Gronlund and Linn (1990, pp. 109–141, 228). The goal of the WATA system can not only support teachers to administer meaningful assessments with the assistance of IT, but can also facilitate teacher education and improve the pre-service teachers’ perspectives about assessment, which is an important training of the PCK (Magnusson et al. 1999, pp. 125–126).

Two empirical studies were conducted after the WATA system was established. The first study was to survey users’ satisfaction of the WATA system, and the results indicated that the in-service teachers were satisfied with most functions of WATA. The second study was to enquire into the potential benefits of implementing the WATA system in teacher education, and the results revealed that the perspectives of pre-service teachers on administering classroom assessment had been improved. It seems that the pre-service teachers become more positive towards the assessment and more familiar with the assessment process with the support of the WATA system.

Table 3. Changes in perspectives of the participants on purpose of assessment by using the WATA system (n = 30)

<table>
<thead>
<tr>
<th>Purpose of assessment</th>
<th>Before WATA training (%)</th>
<th>After WATA training (%)</th>
<th>Cochran’s Q test statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>To distinguish learning outcomes of students</td>
<td>73.33</td>
<td>86.67</td>
<td>2.00</td>
</tr>
<tr>
<td>To understand students’ popular confused conceptions (or misconceptions) about subject matter</td>
<td>70.00</td>
<td>76.67</td>
<td>0.40</td>
</tr>
<tr>
<td>To analyse an individual student’s misconceptions</td>
<td>23.33</td>
<td>70.00</td>
<td>10.89**</td>
</tr>
<tr>
<td>To analyse and improve a teacher’s instructional strategy</td>
<td>10.00</td>
<td>63.33</td>
<td>14.22**</td>
</tr>
<tr>
<td>To realise strength and weakness of test items</td>
<td>20.00</td>
<td>76.67</td>
<td>13.76**</td>
</tr>
<tr>
<td>To realise strength and weakness on choices of a multiple-choice item</td>
<td>16.67</td>
<td>90.00</td>
<td>22.00**</td>
</tr>
<tr>
<td>To realise strength and weakness of a test</td>
<td>26.67</td>
<td>80.00</td>
<td>12.80**</td>
</tr>
</tbody>
</table>

*Cronbach α: 0.65; †Cronbach α: 0.89; **P<0.01.
However, there are some limitations in this research. The evaluation of the WATA system was not completed because the changes of pre-service teachers’ concepts and skills on assessment are neglected in this study. Furthermore, because of the limitation of sample size, we did not have an opportunity to carry out true experimental design. These issues should be considered in further research to make the results more comprehensive. Besides, all participants were limited to pre-service teachers whose subject area was biology; further studies are needed to cover samples with other subject areas and other levels of teachers.

In summary, we have obtained a preliminary result about the potential benefits of the WATA system on a teacher-training program. As a consequence, we suggest that the structure of the Triple-A Model can be considered as a model for Web-based assessment. Moreover, we also believe that the use of Web-based assessment can facilitate teacher training. The research and development in the area of Web-based assessment should be continued to find its applications to teacher education and teaching. Moreover, infusing Web-based assessment, such as Web-based formative assessment, into Web-based instructional environments can improve learners’ learning achievement is also another important issue in a further study.

Acknowledgements

This study was supported by the National Science Council Taiwan, ROC, under Grant No. NSC 91-2511-S-018-024 and NSC 92-2511-S-018 -009. We deeply appreciate their financial support and encouragement. We are grateful for the insightful comments from the referees. We thank Dr Chih-Chiang Yang, Associate Professor of the Overseas Chinese Institute of Technology in Taiwan, for his assistance in the statistical analyses of this study.

References


Table 4. Changes in perspectives of the participants on steps of assessment by using the WATA system (n = 30)

<table>
<thead>
<tr>
<th>Basic steps of assessment</th>
<th>Before WATA training (%)</th>
<th>After WATA training (%)</th>
<th>Cochran’s Q test statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Determining the purpose of testing</td>
<td>60.00</td>
<td>93.33</td>
<td>11.00**</td>
</tr>
<tr>
<td>Constructing the two-way chart</td>
<td>23.33</td>
<td>96.67</td>
<td>21.00**</td>
</tr>
<tr>
<td>Selecting appropriate items according to the two-way chart</td>
<td>23.33</td>
<td>93.33</td>
<td>22.00**</td>
</tr>
<tr>
<td>Preparing relevant items</td>
<td>60.00</td>
<td>100</td>
<td>11.00**</td>
</tr>
<tr>
<td>Assembling the test</td>
<td>100</td>
<td>100</td>
<td>N/A</td>
</tr>
<tr>
<td>Administering the test</td>
<td>100</td>
<td>100</td>
<td>N/A</td>
</tr>
<tr>
<td>Scoring the test</td>
<td>100</td>
<td>100</td>
<td>N/A</td>
</tr>
<tr>
<td>Appraising the test</td>
<td>23.33</td>
<td>100</td>
<td>23.00**</td>
</tr>
</tbody>
</table>

*Cronbach α: 0.73; †Cronbach α: 0.92; **P < 0.01.

N/A = not available.


