

C6: 数码化测验与评量(Digital Assessment & Evaluation)

Technology Use and Student Outcomes: An Empirical Examination of Research Approaches	522
Jing Lei, Qiu Wang	
網路化學習歷程檔案學習者自評之信效度	531
張基成、吳明芳	
新詩賞析平台之研製	542
楊哲青、曾憲雄、王碧玲、楊智凱、翁瑞鋒、蘇俊銘	
A Review of the Strategies for Output Correctness Determination in Automated Assessment of Student Programs	551
Chung Man Tang, Yuen Tak Yu, Chung Keung Poon	
以 QTI 為基礎之線上動態評量管理系統發展及實驗	559
賴阿福、吳明行、陳志鴻	
電腦化動作技能測驗系統之發展與驗證	563
蕭顯勝、宋曜廷、林建佑、邱敬尊	
學習歷程檔案評量研究之發展與趨勢分析	567
劉力君、劉旨峰	
線上閱讀動機量表編製	571
賴怡君、張瑜芳、劉旨峰	
國小五年級數學領域概數與估算單元數位個別指導模式之研發	575
許天維、郭伯臣、劉育隆	
改良式選擇題題型之作文能力測驗方法研究	579
梁惠玲、孫劍秋、吳偉賢、楊志強	
建置攝影課程作品線上評量系統	581
吳振宏、蘇彥寧、歐陽閻	
Reducing the Impact of Inappropriate Items on Reviewable CAT	583
Yung-Chin Yen, Rong-Guey Ho, Li-Ju Chen, and Wen-Wei, Liao	
基於本體論的形成性評量應用於戶外無所不在唐詩教學之成效研究	585
時文中、曾憲雄	

Technology Use and Student Outcomes: An Empirical Examination of Research Approaches

Jing Lei, Qiu Wang
Syracuse University, USA
jlei@syr.edu

Abstract: *The authors argue that to examine the relationship between technology use and student outcomes, the quality of technology use—how and what technology is used—is a more significant factor than the quantity of technology use—how much technology is used. This argument was exemplified by an empirical study that used both angles to examine the association between technology use and student outcomes. When only the quantity of technology use was examined, no significant association was observed. However, when the quality of technology was examined by investigating the specific types of technology uses, significant association was identified between technology use and all student outcomes. Furthermore, different types of technology use showed different influences on specific student outcomes. General technology uses were positively associated with student technology proficiency, while Subject-specific technology uses were negatively associated with student technology proficiency. Social-communication technology uses were significant positively associated with developmental outcomes such as self-esteem and positive attitude toward school. Entertainment/exploration technology use showed significant positive association with student learning habits. None of these technology uses had significant influence on student academic outcome. Specific suggestions for integrating technology into schools and future research were provided.*

Keywords: technology research, quantity vs. quality, technology use, student outcomes

1. Introduction

In the last two decades, generous investments have been made in educational technology around the world. For example, the United States had invested more than \$66 billion in school technology in just 10 years (Quality Education Data, 2004). By 2004, China had spent 100 billion Yuan (about \$13.2 billion) on educational technology (Zhao, 2005), and the annual expense on educational technology was projected to reach 35.5 billion Yuan in 2007 (Okokok Report, 2005). The generous investments were supported by the strongly held premise that technology can help students learn more efficiently and effectively, and as a result increase student academic achievement.

However, this premise on the crucial role of technology in student achievement has not been substantially supported by empirical evidence. In fact, findings from different empirical studies focusing on the effect of technology on learning have been inconsistent and contradictory. On the one hand, some studies have identified significant positive impact of technology use on student outcomes in academic areas such as literacy development (Tracey & Young, 2006), reading comprehension and vocabulary (Scrase, 1998; Stone, 1996), writing (Nix, 1998), mathematics (MacIver, Balfanz & Plank, 1999), and science (Harmer & Cates, 2007; Reid-Griffin, 2003). In addition, positive impacts have been identified in student developmental areas including attitude toward learning and self-esteem (Nguyen, Hsieh & Allen, 2006; Sivin-Kachala & Bialo, 2000), motivation, attendance and discipline (e.g. Mathew, 1997; Sheehy et al., 2005; Twining et al., 2006).

On the other hand, several other researchers argue that technology use may not have any positive impact on student outcomes. For example, in March 2007, the Institute of Education Sciences (IES) released an influential report titled *Effectiveness of Reading and Mathematics Software Products: Findings from the First Student Cohort*. This study, intended to assess the effects of 16 computer software products designed to teach first and fourth grade reading and sixth grade math, using a rigorous random assignment design, found that “test scores in treatment

classrooms that were randomly assigned to use products did not differ from test scores in control classrooms by statistically significant margins” (Dynarski, et al., 2007, p.xiii).

Furthermore, some studies suggest that technology use might even harm children and their learning (e.g., Healy, 1998; Stoll, 1999). A study of the Trends in International Mathematics and Science Study (TIMSS) reported that technology use was negatively related to science achievement amongst eighth graders in Turkey (Aypay, Erdogan, & Sozer, 2007). Another TIMSS study found that extensive use was related to lower science scores (Antonijevic, 2007). Similarly, based on data collected from 175,000 fifteen-year-old students in thirty-one countries, researchers at the University of Munich announced that performance in math and reading had suffered significantly among students who had more than one computer at home (MacDonald, 2004).

Existing research on the relationship technology on student learning presents a mixed message (Andrews, 2007; O'Dwyer, et.al, 2005; Torgerson & Zhu, 2003). Such mixed and often conflicting findings make it difficult to draw conclusions about the effects of technology, to provide meaningful advice to those who make decisions about technology investment in education, and to make practical suggestions for integrating technology into schools.

There are at least two problems contributing to the controversy over the relationship between technology use and student outcomes. The first is that technology is often examined at a very general level (Zhao, 2003). Many studies “treat technology as an undifferentiated characteristic of schools and classrooms. No distinction is made between different types of technology programs” (Wenglinsky, 1998, p.3). We know that technology is a very broad term that includes many kinds of hardware and software. These technologies may have different impacts on student outcomes. Even the same technology can be used differently in various contexts to solve all kinds of problems (Zhao, 2003), and thus have “different meanings in different settings” (Peyton & Bruce, 1993, p.10). Treating technology as if it is a single thing obscures the unique characteristics of different technologies and their uses.

The second problem is the focus of the studies. Most studies focus on the impact of the quantity of technology use, in other words, *how much* or how frequently technology is used, but ignore the quality of technology use, that is, *how* technology is used. For example, many studies examine the relationship between how much time students spend on using computers or how often they use computers and their achievement (e.g., Du, Havard, Yu, & Adams, 2004; Mann, Shakeshaft & Becker, 1999). However, research suggests that the quality of technology use is more critical to student outcomes than the quantity (Burbules & Callister, 2000; Lei & Zhao, 2007; McFarlane, 1997). Thus, the necessary next step is to examine how different uses of diverse technologies affect student learning.

This study investigates the relationship between technology use and student outcomes by comparing the association between the quantity of technology and student outcomes to the association between the quality of technology use and student outcomes. This approach differs from many previous studies in at least two aspects. First, it studies technology at a more specific level instead of technology in general. Second, to better discern the quality of technology use, this study focuses on different uses of technology rather than on specific technological objects such as hardware or software. “Technology use” is the application of a technology function to solve practical problems (Zhao, 2003). The focus is technology-in-context. Examining technologies from this angle allows us to discern the different uses of the same technologies so that the nature of different technology uses can be better understood.

2.Methods

Participants were 7th and 8th grade students and teachers in a northwestern middle school in the United States. This was a comparatively small school, with a total enrollment of 237 for two grades, and the student-teacher ratio was 9.1. This school had rich technology resources such as one-to-one laptops. Data were collected through surveys and interviews. Student GPAs were collected from their school records.

Survey

The survey was administered twice, at the beginning and the end of the academic year. It included four sections. The first section asked about demographic information including SES, grade, and gender. The second section measured students' information technology proficiency. The third section, student outcomes, included learning habits and developmental outcomes. Developmental outcomes included self-esteem, attitude toward schooling, and social skills. Questions in this section were Likert scale questions measured on a scale of 1-5 with 1 indicating "strongly disagree" and 5 indicating "strongly agree". Based on data collected from interviews, the fourth section, technology usage, listed twenty-eight specific technology uses ranging from emailing and using PowerPoint for presentation to playing games and creating websites. Participants were asked to rate how often they worked with each of these technology uses. Academic achievement information (GPA) was collected from student report cards.

The survey was administered to all 237 students in this school. Among them, 207 students returned the first survey, 208 students returned the second survey, and 177 students filled out both surveys. Data from students with more than one third of all responses missing were deleted (N=34), and data from special education students were deleted (N=10) because the only technologies they used were assistive technologies, which were not included in this study. Therefore, altogether 133 students' data were retained for final data analysis. Of the 133 students, sixty-four (48%) were male, sixty-nine (52%) were female, sixty-four (48%) were 7th graders, and 69 (52%) were 8th graders.

Data Analysis

Reliability Check. Reliability was checked for researcher designed scales. The reliability of the student Technology Proficiency Scale, Learning Habit Scale, and Developmental Outcome Scale was 0.77, 0.77 and 0.90, respectively.

Categorizing Technology Use. The twenty-eight specific technology uses in section four of the survey were categorized into five groups according to the purposes and nature of use. Subject-specific technology uses included technology uses related to subject content. Social/communication technology uses included technology usage for communication or social interaction purposes. Construction technology uses included technology uses that created products. Entertainment/exploration technology uses included technology uses for fun and self-interest. General technology uses included technology that can be applied to any content area and for any purpose.

Linear Regression analyses were conducted to examine the relationships between technology uses and student outcomes. Interview data were coded and analyzed according to specific research questions.

3. Results

This section first examines the relationship between the quantity of technology use—how much time was spent on computers on student outcomes, and then examines the relationship between the quality of technology use—how technology was used and student outcomes.

The quantity of technology use and student outcomes

Time spent on computers everyday

Descriptive analysis results revealed that, as shown in Table 2, 32.3% of the students spent less than two hours a day on computers, 30.8% of the students spent about two to three hours a day on computers, and 36.9% of them spent more than three hours a day on computers.

Table 2: Time Spent on Computers

Time	Percent of Students
Less than one hour	7.7%
About 1-2 hours	24.6%
About 2-3 hours	30.8%
More than 3 hours	36.9%

The relationship between the quantity of technology use on student outcomes.

To examine the relationship between the quantity of technology use and student outcomes, regression analysis was conducted to analyze how the time spent on computers affect each of the four student outcomes: GPA, technology proficiency, learning habits, and developmental outcomes.

Table 4: Relationship between the quantity of technology use on student GPA

	B	Std. Error	t	p
(Constant)	.279	.177	1.58	.120
how much time do you spend on computers everyday?	-.074	.052	-1.41	.164

Table 4 shows that how much time spent on computers everyday was not significantly associated with student GPA ($P = .12$). In other words, no strong association was identified between the quantity of technology use and student GPA.

Similarly, regression analyses revealed non-significant relationship between the quantity of technology use and student technology proficiency ($R^2 = .01$, $B = 0.2$, $p = 0.27$), learning habits ($R^2 = .01$, $B = 0.06$, $p = 0.26$), and developmental outcomes ($R^2 = .01$, $B = 0.08$, $p = 0.20$).

In summary, when only looking at the quantity of technology use, data analyses revealed no significant relationship between technology use on student outcomes.

The quality of technology use and its relationship with student outcomes

This section examines the relationship between technology use and student outcomes from another angle: the quality of technology use, that is, how technology was being used. Specifically, regression analyses were conducted to examine if students outcomes were affect by the five types of technology uses: general technology use, subject-specific technology use, social-communication technology use, construction technology use, and entertainment/exploration technology use.

Relationship between different technology uses and student academic achievement

Table 5 presents the results of regression analysis on the relationship between different technology uses and student academic achievement, as represented by student GPA. Effect sizes are also included to show the strength of the relationship.

Table 5: Relationsihp Between Technology Uses and Student GPA

Effect	Regression coefficient			t	p
	β	SE(β)	Effect size		
(Constant)	5.735	2.243		2.56	.012
General tech use	.092	.159	0.10	.58	.565
Subject-Specific tech use	.023	.103	0.04	.22	.828
Social-Communication tech use	.120	.099	0.21	1.21	.230
Construction tech use	.002	.098	0.00	.02	.982
Entertain/Explore tech use	-.129	.095	-0.24	-1.36	.177

As shown in Table 5, using technology for social-communication purposes had some positive influence on student GPA. Although this influence was not statistically significant, an effect size of 0.21 on GPA was noteworthy compared with a possible 0.33-0.50 effect size gain on student performance based on “everything that happens to a student” (Kane, 2004, p.3) across one academic year. This association was probably the result of student using social-communication technologies to communicate with teachers regarding assignments and questions on lectures. With these means of communication, students could receive responses more quickly than with traditional methods.

Social-communication technologies also provided students more opportunities and avenues to ask questions. A number of teachers mentioned that they often received e-mail messages from students who were too shy to ask questions in the classroom. Students also mentioned emailing their teachers during the interviews. They reported that it was easier and more convenient to ask questions or set appointments with teachers through e-mail.

Entertainment-exploration technology uses were a negative associated with student GPA ($ES = -0.24$, $p > .05$). This was likely the result of using study time for entertainment. The more time spent on these technology uses, then the less time left for learning. In the interview, students talked about “other students” who spent too much time playing computer games and commented that that was not good for their learning.

It should be acknowledged, however, that the relationships between student GPA and technology uses identified in this study were not necessarily causal, but associative in nature. This applies to other relationships identified through regression analysis in this study.

Relationship between technology uses and student technology proficiency

A regression analysis was conducted to examine the relationship between different technology uses and student technology proficiency. As shown in the following two tables, this regression model was statistically significant ($P < .05$) and it explained 14.3% of the total variation.

Table 7: Relationship between Technology Use and Student Technology Proficiency

Effect	Regression coefficient			t	p
	β	SE(β)	Effect size		
(Constant)	8.051	1.995		4.036	.000
General tech use	1.416	.774	0.32	1.829	.071
Subject-Specific tech use	-1.291	.533	-0.43	-2.422	.017
Social-Communication tech use	-.280	.504	-0.10	-.555	.580
Construction tech use	-.596	.495	-0.21	-1.206	.231
Entertain/Explore tech use	.322	.449	0.13	.718	.474

As shown in Table 7, general technology use had a marginally significant influence on student technology proficiency ($t=1.83$, $P = .07$), while subject-specific technology use had a significantly negative influence on student technology proficiency ($t= -2.42$, $P < .05$). This is understandable in that when using more general technologies, the tasks are not certain, the technologies vary, students often have to explore new features of certain technologies, and thus have the opportunity to learn more about technology. However, when they use subject-specific technologies to learn, the tasks are focused on specific subject content, and the procedures to accomplish the tasks are generally similar. Therefore, once students know how to follow these procedures there are no more technological challenges and no opportunities to expand technology knowledge and skills.

Relationship between technology uses and student outcomes:

The relationship between technology use and student learning habits and developmental outcomes was also examined by using regression analysis. To better compare the relationships between different technology uses on different student outcomes, the following table summarizes the overall findings:

Table 8: Relationships between Technology Uses on Student Outcomes (Effect Size)

Student Outcomes	GPA	Technology Proficiency	Learning Habits	Developmental Outcomes
Student Tech Uses				
General tech use	0.10	0.32	0.11	-0.03
Subject-Specific tech use	0.04	-0.43*	0.04	-0.03

Social-Communication tech use	0.21	-0.10	-0.09	0.35*
Construction tech use	0.00	-0.21	-0.16	0.00
Entertainment/Exploration tech use	-0.24	0.13	0.51**	-0.08

Note: *: $0.01 < p < 0.05$ **: $p < 0.01$

Table 8 lists the effect size of the regression coefficient of each technology use on every student outcome. Results in Table 8 show that different technology uses have different influences on specific student outcomes. General technology uses were positively associated student technology proficiency, but the influence on other outcomes was minimal. Subject-specific technology use had a significantly negative association with student technology proficiency. In addition to the noticeable positive association with student academic achievement, social-communication technology use had a significantly positive influence on student developmental outcomes ($ES = 0.35, p < .05$). It is arguable that the more students used technology for social-communication purposes, the more they felt socially connected, a very important feeling for teenage students who need support from their peers and adults (Wighting, 2006).

Entertainment-exploration technology use significantly influenced student learning habits ($ES = 0.51, P < .01$). In the interview, students reported that entertaining activities and exploring with technology could help them organize their learning tasks better. For example, remembering and following rules in computer games may help students follow instructions in classrooms more efficiently, and ease in following directions should be beneficial to students' attaining learning outcomes. However, it seems this potential advantage was nullified or even outweighed by the consequences of spending too much time on entertainment-exploration technology use.

4. Conclusions and Implications

This study investigated the relationship between technology use and student outcomes by examining both the quantity of technology use—how much time was spent on computers and the quality of technology use—how technology was used. When only examining how much time was spent on computers, no significant relationship was found between technology use and any student outcomes. However, when how technology was used was examined, significant association was identified between technology use and most student outcomes. General technology use helped improve student technology proficiency, while subject-specific technology use significantly impeded the development of technology proficiency. Social-communication technology uses had a significant positive association with student developmental outcomes and a moderate positive association with student academic achievement. Furthermore, the same type of technology uses had different influences on different student outcomes. For example, entertainment-exploration technology uses helped improve student learning habits. However, it might affect student academic achievement if too much time is spent on using technology for entertainment.

Findings from this study have significant implications for policy-making, research and practices regarding technology integration in schools. (1) *Focusing the quality of technology use*. Results from this study suggest that technology can have significant influence on student outcomes, but the influence depends not only on how the technology was used, but also on how the influence was examined and measured. This calls for more emphasis on the quality of technology use in both research and practices. For technology to have meaningful impact on teaching and learning, close attention must be paid on the quality of technology use: how is it being used, what is used, and for what purposes. (2) *Be realistic about the impact of technology*. Results from this study show that student academic outcome (GPA) may not be easily improved through the use of technology. This is probably because student performance, especially academic outcomes measured by GPA, is influenced by many factors. Technology usage is just one of these factors. The association between technology use and student outcomes is not determined merely by the particular technology uses, but is mediated by environmental factors, the users, the technology, and the constantly

changing interactions and mutual influences. Therefore, it may be unrealistic to expect dramatic changes in student performance through one or two specific technology uses. (3) *Set specific educational goals for technology use.* Because different technology uses have different influences on student outcomes, to facilitate technology use in schools and to accurately assess the effectiveness of specific technology uses, it is important to set clear educational goals even before technologies are purchased and installed.

Suggestions for future research: Additional research needs to be conducted along the following lines. First, to identify more effective technology uses. Technology has the potential to improve teaching and learning. However, for this potential to be realized, it must be “properly used”. More research on effective technology usage is required to help policy-makers, educators and practitioners understand what technology uses are “proper use” of technology so that the potential benefits of technology can be reaped by teachers and students alike. Second, the effectiveness of technology use is contingent on the specific student outcomes. Academic achievement should not be the only criterion for evaluating the meaningfulness or effectiveness of technology use. Some other outcomes are also important components of school education including student behavior, attitude, self-esteem, digital literacy, and career aspiration. Exploration of these aspects can help enhance the effectiveness of using technology to help to develop complete learners. Third, there is a need to explore and develop evaluation methods and instruments that evaluate student learning with technology. Student technology use and learning is experience-related and at times hidden or subtle; consequently, it cannot be assessed through traditional outcome evaluation. Some alternative assessment methods such as performance assessment, essays and portfolios might be more effective in assessing student learning with and about technology. Fourth, this study examined the quality of technology use by looking at five types of technology use. This categorization was helpful in this study to reveal the critical difference between the quantity and the quality of technology use, but it also had some limitations. For example, since each type of technology uses was consisted of several specific technology uses, the grouping might have canceled out the differences among these specific technology uses. Research needs to further explore effective ways to measure the quality of technology use.

Reference:

- Andrews, R., Freeman, A., Hou, D., McGuinn, N., Robinson, A. & Zhu, J. (2007). The effectiveness of information and communication technology on the learning of written English for 5- to 16-year-olds. *British Journal of Educational Technology*, 38(2), 325-336.
- Antonićević, R. (2007). Usage of Computers and Calculators and Students' Achievement: Results from TIMSS 2003. Paper presented at the International Conference on Informatics, Educational Technology and New Media in Education, Sombor, Serbia, Mar 31-Apr 1, 2007.
- Aypay, A., Erdogan, M., & Sozer, M. A. (2007). Variation among schools on classroom practices in science-based on TIMSS-1999 in Turkey. *Journal of Research in Science Teaching*, 44 (10), 1417-1435.
- Burbules, N., & Callister, Jr., T. (2000). *Watch IT: The promises and risks of new information technologies for education*. Boulder, Colorado: Westview Press.
- Du, J., Havard, B., Yu, C., & Adams, J. (2004). The impact of technology use on low-income and minority students' academic achievement: Educational longitudinal study of 2002. *Journal of Educational Research & Policy Studies* 4(2), 21-38.
- Dynarski, M., Agodini, R., Heaviside, S. N. T., Carey, N., Campuzano, L., Means, B., et al. (2007). *Effectiveness of Reading and Mathematics Software Products: Findings from the First Student Cohort*. Report to Congress. Washington D. C.: Institute of Education Sciences.
- Harmer, A.J., & Cates, W. M. (2007). *Designing for Learner Engagement in Middle School Science: Technology*,

- Inquiry, and the Hierarchies of Engagement. *Computers in the Schools*, 24(1-2),105-124.
- Lei, J., & Zhao, Y. (2007). Computer Uses and Student Achievement: A longitudinal Study. *Computers & Education*, 49(2), 284-296.
- MacIver, D. J., Balfanz, R., & Plank, S. B. (1999). An "elective replacement" approach to providing extra help in math: The Talent Development Middle Schools' Computer- and Team-Assisted Mathematics Acceleration (CATAMA) Program. *Research in Middle Level Education Quarterly*, 22(2), 1-23.
- Mann, D., Shakeshaft, C., Becker, J., & Kottkamp, R. (1999). *West Virginia's Basic Skills/Computer Education Program: An Analysis of Student Achievement*. Santa Monica, CA: Miken Family Foundation.
- McFarlane, A. (1997). What are we and how did we get here? In McFarlane, A. (Ed.) *Information technology and authentic learning: realizing the potential of computers in the primary classroom*. Routledge.
- Nguyen, D. M., Hsieh, Y. J., Allen, G. (2006). The Impact of Web-Based Assessment and Practice on Students' Mathematics Learning Attitudes. *The Journal of Computers in Mathematics and Science Teaching* 25(3), 251-279.
- Nix, C.A.G. (1998). The impact of e-mail use on fourth graders' writing skills. *Dissertation Abstracts International*, 60/03-A (Order No. AAD99-21889).
- O'Dwyer, L. M., Russell, M., Bebell, D., & Tucker-Seeley, K. R. (2005). Examining the relationship between home and school computer use and students' English/language arts test scores. *Journal of Technology, Learning, and Assessment*, 3(3). Available from <http://www.jtla.org>
- Okokok Report, (2004). "2005 Market Report of China's Educational Technology Development and IT Application Trends". <http://www.okokok.com.cn/Shop/Class41/200411/167.html> (accessed August 23, 2007).
- Peyton, J. K., & Bruce, B. C. (1993). Understanding the multiple threads of network-based classrooms. In Bruce, B. C., Peyton, J.K., & Batson, T. (ed.) *Network-Based Classrooms: Promises and realities*(pp. 9-32), Cambridge, England: Cambridge University Press.
- Quality Education Data (QED). (2004). 2003-2004 Technology Purchasing Forecast <http://www.qeddata.com/marketkno/researchreports/techpurchaseforecast.aspx>
- Reid-Griffin, A. (2003). Technology: A Tool for Science Learning. *Meridian: A. Middle School Technologies Journal*. 10(2). Retrieved February 17, 2009 from <http://ncsu.edu/meridian/sum2003/science/index.html>
- Scrase, R. (1998). An Evaluation of Multi-sensory speaking-computer bases system (Starcross-IDL) designed to teach the literacy skills of reading and spelling. *British Journal of Educational Technology*, 29 (3), 221-224.
- Sheehy, K; Kukulska-Hulme, A; Twining, P; Evans, D; Cook, D and Jelfs, A., with Ralston, J; Selwood, I; Jones, A; Heppell, S; Scanlon, E; Underwood, J and McAndrew, P,(2005) Tablet PCs in schools: A review of literature and selected projects, Becta Research, Retrieved 11/09/2006 from http://www.becta.org.uk/corporate/publications/documents/tablet_pc.pdf.pdf
- Sivin-Kachala, J., & Bialo, E. (2000). *2000 Research report on the effectiveness of technology in schools* (7th ed.). Washington, DC: Software and Information Industry Association.
- Stone, T. T., III (1996). The academic impact of classroom computer usage upon middle-class primary grade level elementary school children. *Dissertation Abstracts International*, 57/06-A (Order No. AAD96-33809).
- Torgerson, C. & Zhu, D. (2003). A systematic review and meta-analysis of the effectiveness of ICT on literacy learning in English, 5–16. In Research evidence in education library. EPPI-Centre, Social Science Research Unit, Institute of Education, London. Retrieved September 1, 2006, from <http://eppi.ioe.ac.uk/reel>
- Tracey, D.H., & Young, J. (2006). Technology and early literacy: the impact of an integrated learning system on high risk kindergartners' achievement. ERIC ED491554.
- Twining, P; Evans, D; Cook,D; Ralston, J; Selwood, I; Jones, A; Underwood, J; Dillon, G., Scanlon, E; Heppell, S;

- Kukulska-Hulme, A; McAndrew, P., & Sheehy, K. (2006). Tablet PCs in Schools: Case study report, Becta Research, Retrieved 11/09/2006 from:
http://www.becta.org.uk/corporate/publications/documents/tabletpc_report.pdf
- Wenglinsky, H. (1998). *Does it Compute: the relationship between educational technology and student achievement in mathematics*. Policy Information Center Report. ETS. ED425191.
- Zhao, G., (2005) A report on the status quo of Chinese education technology integration. *China Education and Research Network*. Retrieved August 10, 2007, from
http://www.edu.cn/li_lun_yj_1652/20060323/t20060323_126669.shtml
- Zhao, Y. (2003). What Teachers Need to know about Technology? Framing the question. In Zhao, Y. (ed). *What Should Teachers Know about Technology: Perspectives and Practices* (pp.), Greenwich, CT: Information Age Publishing.

網路化學習歷程檔案學習者自評之信效度

Reliability and Validity of Learner Web-Based Portfolio Self-Assessment

張基成

台灣師範大學科技應用與人力資源發展系
電子郵件：amchang@ntnu.edu.tw

吳明芳

台北市立松山家商資料處理科
電子郵件：wumingfang@ms26.url.com.tw

【摘要】本研究探討網路化檔案學習者自評的信、效度。研究樣本為某高職修習「計算機應用」課程的三十六位學生。學生藉由網路化檔案評量系統進行個人檔案的製作、觀摩與自評。結果顯示：(1)學習者兩次檔案自評結果具有高度的一致性。(2)學習者自評與教學者評結果具高度一致性且無顯著差異。(3)學習者檔案自評分數與測驗成績結果具高度一致性，顯示網路化檔案自評可以反應學習成就。簡言之，網路化檔案自評兼具信、效度。

【關鍵詞】學習歷程檔案、網路化學習歷程檔案、網路化檔案評量、自評

Abstract: This study examines the reliability and validity of Web-based portfolio self-assessment. The research samples consists of thirty-six students who take "Computer Application" course of the second grade in some vocational high school. The students use the Web-based portfolio assessment system for creating their portfolio, emulating peers' portfolio, and performing self-assessment. The results demonstrate that there exists a high consistence between the two self-assessment. There exists a high consistence and no significant difference between student self-assessment and teacher assessment. The consistence between peer-assessment and test score is high, meaning the Web-based portfolio self-assessment may reflect learning achievement. In short, Web-based portfolio self-assessment has adequate reliable and validity.

Keywords: Portfolio, Web-Based Portfolio, Web-Based Portfolio Assessment, Self-Assessment

1. 研究背景

學習歷程檔案 (learning portfolio) 乃學習者經過一段時間，有目的、有組織地蒐集其學習過程、進步紀錄、心得感想、反省資料、學習成果等資料，可供檢視學習者學習成就。Russell 與 Butcher (1999)、Springfield (2001) 指出，傳統紙本式學習歷程檔案會因為長時間進行資料蒐集以致數量增加，造成展示與觀摩上的不便。王等元 (2000)、岳修平、王郁青 (2000) 不約而同的表示學習歷程檔案 (learning portfolio) 的網路化將是未來的應用趨勢；Yancey (2001b) 在從事網路化學習歷程檔案的研究之後亦提出這樣的看法。Chang (2001)、Chang & Tseng (2009a)。指出藉由網路上學習檔案的分享，不但能幫助教師了解學生的學習狀況，更有助於師生彼此間互動的增進；同時對學生而言，亦能了解同儕間的學習成果及過程。傳統紙本式學習歷程檔案在儲存空間及管理上的問題是可以藉由網路技術加以克服。

檔案評量 (portfolio assessment) 乃依據特定目的有系統地將被評量者在學習過程與結果等相關紀錄集結成冊，藉以評定學習者的努力、成長、進步與成就情形的一種評量方式 (Chang, 2008)。簡言之，檔案評量即學習歷程檔案用於評量學習者學習成就。網路化檔案評量 (Web-based portfolio assessment) 指的是透過網路來輔助檔案評量的進行，評量者可於線上進行學習者自評、同儕互評、教學者評量或其它相關評量等活動 (Chang & Tseng, 2009b)。Gadbury-Amyot 等人 (2003) 認為，在課程當中讓學生進行自評活動可以提升大學生的學業能力，而這些自評活動可以透過檔案評量來進行。通常在網路化檔案評量過程中，常會要求學生針對自己的檔案在網路上做自評。更進一步來說，應是以自己的檔案內所呈現的學習過程與結果資料，對自己的學習在線上做出評論與反省，包括學習成果是否有達到預期目標、是否有進步、有哪些地方需要改善等。目前，許多研究較常探討同儕互評議題，較少討論學習者自評的議題。尤其是檔案評量過程中自評的信、效度問題，更少討論。Oskey、Schallies 與 Morgil (2008) 綜合幾個檔案評量信、效度的研究結果，結論認為檔案評量是一個恰當且可靠的評量方法，但可惜的是並未說明是否也包括學生自評。如何進行自評？自評的信度如何？自評的效度如何？這些都是值得探討的議題。

根據上述研究背景，本研究之目的為探討網路化學習歷程檔案學習者自評的信、效度。相對

應的研究問題與統計方法如表 1。

表 1 研究目的、研究問題與統計方法對照表

研究目的	研究問題	進行方式	統計方法
學習者自評的信度	學習者兩次自評結果是否一致？	再測一致性	Pearson 積差相關
學習者自評的效度	學習者自評與教學者評量結果是否一致？是否有顯著差異？	外在效標關聯效度（效標為授課教師所評分數）	Pearson 積差相關/t test
	學習者自評分數是否與測驗成績一致？自評分數（扣掉作品）是否與作品成績一致？		Pearson 積差相關

2. 文獻探討

2.1 網路化檔案評量

岳修平、王郁青（2000）在其研究中表示，採用學習歷程檔案（learning portfolio）使學生對自己所學與評量具所有權威，並對自己的作品產生責任感，透過其內容的真實性讓學生檢驗他們的知識與技能。張麗麗（2002）指出透過檔案可以評量學習過程與結果，並提供認知與情意並重及尊重學生個別差異的檢測。學習歷程檔案是一種可以真實評估出學習者能力的評量方式，透過學習者自行組織檔案內容可以真正檢視到學習者的學習成效。Hult（2001）表示學習歷程檔案可以讓學生自己明確發現及確認自己的問題。就後設認知理論中的自我監控（self-monitored）及自我調整（self-regulated）策略而言，具有自我監控能力的人會不斷的注意自己學習進展的情形，分辨較佳與較差之處，在學習中不斷地發現及澄清問題（王等元，2000）。Luca 與 McMahon（2006）將網路化學習檔案（Web-based portfolio）做為線上學習的自我監督與評量工具。Chang & Tseng（2009b）的研究證實網路化檔案評量的使用對學生的自評能力有提昇作用。綜上所述，在網路化檔案評量環境下，學習者線上自我檢視與評量扮演重要角色。

2.2 自評信效度

2.2.1 自評信度

陳英豪、吳裕益（1991）指出信度（reliability）是評量結果的一致性，即評量結果可不可靠。余民寧（2003）認為信度是多次評量結果間的一致性（consistency）或穩定性（stability）。陳嘉鴻（2002）指信度是相同的人在不同的時間，以相同的評量所得結果的一致性；如果兩次評量結果一致，表示結果具有穩定性、可靠性或可預測。王家玲（2002）指信度是評量結果具有一致性、穩定性的程度。對於信度的檢驗，一般多以內部一致性及再測信度來表示評量信度的高低。評量分數要具有意義，它必須具有相當程度的一致性與可靠性（張麗麗，2002）。

針對自評信度，研究者提出如下定義：是對於同一份評量採用複本再測，或者在不同時間與情境下施測結果的一致、穩定及可靠程度。對於信度的檢驗方法上，可以採用皮爾森積差相關（Pearson product-moment correlation）找出兩次評量結果的相關係數（余民寧，2003；張麗麗，2002；劉旨峰、林珊如、袁賢銘，2002；Derham & Diperna, 2007；Gadbury-Amyot et al., 2003；Lenze, 2003；Liu, Lin, Chiu, & Yuan, 2001；Lin, Liu, & Yuan, 2001）。評量表的內部一致性可用 Cronbach's α 係數來檢測（余民寧，2003；Derham & Diperna, 2007）。

2.2.2 自評效度

陳英豪、吳裕益（1991）指出效度（validity）是一個評量能夠測量到它所欲測量的特質或功能的程度，即評量結果正不正確。余民寧（2003）表示效度是評量分數的有效程度，亦即評量能夠提供適切資料以作決策的程度。陳嘉鴻（2002）表示效度是評量分數的正確性；簡言之，就是指一個評量能夠測量到他所想要測量的特質的程度。針對自評效度，研究者提出如下定義：是指一個評量能夠測量到它所欲測量的特質或功能的程度，亦可稱為某一評量能提供適切資料以作成決策參考的程度。

通常若欲檢驗學習者自評的效度，可用授課教師的評分或助教學理的評分作為校標以茲參照（Sluijsmans, Dochy, & Moerkerke, 1999）。余民寧（2003）表示，若以外在效度來解釋評量的效度，需要一個外在效標以供參考，而授課教師的評分可以視為一個外在效度指標。Gadbury-Amyot

et.al (2003) 的研究是以學生測驗分數當做外在效標，檢驗其與檔案評量分數的一致性。兩者一致性越高，表示檔案評量的效度越高。研究者認為，學習者自評的效度可以運用外在效標的方式予以檢驗，教學者評分與學生測驗分數都可以是外在效標的來源。張麗麗 (2002) 表示為求外在效標具有更高的效度，教學者數量多一些比越好或者接受過評分訓練。研究者認為如果可以的話，再加上對做為外在效標的教學者評分結果之間一致性的檢驗，外在效標的參考程度會更高。

3. 研究方法與步驟

3.1 研究對象

參與實驗者為台灣北部某高級職業學校會計事務科二年級修習「計算機應用」課程的一個班級學生，共四十三人。刪除七位檔案內容不完整足以影響統計結果者（譬如學習目標未設定、作品未繳交、反思撰寫不全等），剩三十六人之檔案作為統計分析用；其中男生 12 名，女生 24 名。該門課學分數為 3 學分，每週上課 3 小時。教學實驗為期十二週，教學內容為「Word 文書處理：版面設定及長文件編輯」（兩單元）。授課教師要求完成的作業是此次實驗中學習者學習歷程檔案內蒐集的主要內容。實驗期間，學習者使用網路化檔案評量系統進行個人檔案的製作、觀摩與自評，教師則透過系統檢視學習者的檔案並評量其學習表現。

3.2 檔案評量標準發展

根據文獻歸納出初步的網路化檔案評量標準，經與授課教師討論及修正後製成量表。之後委請三位學者專家協助專家效度的建立（包含表面效度與內容效度），針對所給的意見進行修正。評量標準分為檔案製作、學習目標、作品、反思、態度、其它六個面向，共 32 項指標，每個指標依據被評量者所達到的績效水準給予 1、1.5、2、2.5、3、3.5、4、4.5 與 5 的點數。

在學習目標的衡量上，關注學習者個人學習目標的達成程度。作品面向細分為正確性、恰當性、豐富性及完整性、難度、創意與創新等題項，同時也納入作品歷程紀錄。在反思面向中，考核的重點包含學習者對學習目標、作品、學習成就、學習態度等的反省思考。倘若學習者在這些項目中的反思明確指出自己的缺點及可改進的地方，則可獲較高評價；但如果反思內容呈現的是該位學習者學習上的負面缺失，如「過程中自己抱持偷懶的態度進行學習，下一次的學習要改正此缺點，態度要轉為積極。」針對上述情況，本研究顧及學習者因擔心扣分而怯於寫出缺點，因而在評分時並不扣分。

相較於傳統紙面式檔案評量，網路可增添傳統檔案所欠缺的互動性 (Yancey, 2001)，因而此評量標準融入了一些網路的特性。在檔案製作面向上，內容豐富程度指標考量了學習者上傳掃描圖檔、Word 檔、PowerPoint 檔、影音檔等狀況；在反思面向上，增加了對觀摩同儕表現的反思，針對他人回饋的反思的指標；在態度面向上，增加了線上觀摩、欣賞、互評與回饋的表現、網路資源的分享、線上互動的數量及品質等評量項目。

3.3 檔案評量標準項目分析

首先針對預試量表進行項目分析 (item analysis)。在鑑別度 (discrimination) 部分，研究者將每一指標平均數分為高(27%)、低分(27%)兩組後，再進行兩組平均數比較 (t 檢定)。結果顯示，每一題 t 檢定的結果 (t 值，即決斷值) 皆達顯著水準，表示每一題的鑑別力足夠，不必刪除。在一致性 (consistency) 部分，研究者將每一指標平均數與所有指標平均數進行 Pearson 積差相關檢定。結果顯示，每一題檢定結果 (Pearson 積差相關係數) 皆達顯著水準，顯示每一指標之間的一致性足夠，不必刪除。

3.4 檔案評量標準效度檢驗

抽取共同因素面向方式是採用主成份分析法，萃取特徵值大於 1 的因素面向。量表預試各面向的抽樣適切性量數 (Kaiser-Meyer-Olkin, KMO) 值皆大於 0.5 (表 2)，可進行因素分析並利用其主成份分析法 (Principal Component Analysis, PCA) 建構效度。因考量各因素面向之間具一定程度之相關性，以最大變異法 (varimax) 進行斜交轉軸 (oblique rotation)；而 Bartlett 球形檢定之近似卡方分配皆達顯著，表示各指標之間有共同因素存在，適合進行因素分析。第一次因素分析結果顯示，「態度」面向內有一指標的因素負荷量小於 0.3，將此指標刪除。第二次因素分析結果顯示，因素負荷量皆大於 0.3，且效度係數皆大於 0.1，故保留所有指標。

最後共萃取五個特徵值大於 1 的因素面向。各面向的累積變異量皆大於 45%，顯示各面向皆具一定之效度。作品面向的累積變異量最大，態度面向最小。

表 2 量表預試因素分析

面向	KMO 值	累積變異量 (%)
檔案製作	0.67	56.26
學習目標	0.84	58.10
作品	0.89	66.71
反思	0.80	50.64
態度	0.69	45.41
其它	無	無
整體	0.86	72.09

註：「其它」面向僅有一題，不做因素分析

3.5 檔案評量標準信度檢驗

預試與正式評量各面向的 Cronbach's α 值皆大於 0.7，屬於高信度。顯示各項指標之間具有高度的內部一致性，因此預試後保留各項指標。

表 3 量表兩次施測 Cronbach's α 值

面向	預試	正式施測
檔案製作	0.91	0.91
學習目標	0.70	0.94
作品	0.77	0.93
反思	0.84	0.90
態度	0.73	0.74
其它	無	無
整體	0.92	0.97

註：「其它」面向僅有單獨一項，無法計算 Cronbach's α 值

3.6 網路化學習歷程檔案評量系統

此網路化檔案評量系統提供檔案內容項目 (entry) 按鈕及輸入資料的表單視窗，使用者只要點選檔案內容項目 (entry) 按鈕及根據表單上的說明輸入資料，即可完成個人檔案的製作。每人的檔案內容項目名稱與檔案內容格式相同，有利於評量者評分。

系統功能(如圖 1 左邊按鈕)主要有：(一)檔案製作指引。(二)檔案製作區(僅供學生使用)：含基本資料、學習目標設定、作品繳交、自我反思、其它等。(三)檔案評量區：學生與教學者身份不同，若為教師進入則為教師評分；若為同學登入則為自評與同儕互評(內含檔案評量指標之等地勾選)。評分的同時，也可以瀏覽被評者的檔案內容。(四)檔案成績區：可查閱學習者自評分數、同儕互評分數、教學者評分數，及觀摩優良作品等。(五)課程說明區：包括課程基本資料及教師的教案。

上方的工具列(如圖 1 上方按鈕)功能有：(一)系統公告：觀看系統最新公告。(二)討論專區：可進行課程相關及檔案製作的討論。(三)基本資料：可查看個人基本資料。(四)網頁式個人檔案瀏覽：顯示每個人的檔案內容。(五)檔案分項瀏覽：依據分類的內容項目顯示所有成員學習檔案中的該項目內容。(六)發送信件：可發送信件給每一位參與者。(七)單元切換：可切換不同的授課單元，讓每個單元呈現不同的檔案內容。



圖 1 自評

自評時，學習者點選系統左邊的「檔案評量區」，系統便會出現所有學者姓名，如圖 1。按自己名字旁邊的「自評」，在畫面中間上面會出現一整列的檔案內容選項（基本資料、學習目標、反思內容、單元作品、其他內容、「評分」、教師回饋、同儕回饋、參與情形等）。點選中間的「評分」按鈕，會另外出現一個網頁視窗評分表。此時，教學者可點選每一項評量指標旁的分數，及指標下面的文字方塊填寫回饋意見。評分時，教學者可同時瀏覽學習者的檔案內容，以做為評分的參考。評分完畢後，可以選擇預覽評分結果、重新評分、或者確定送出，如圖 2。

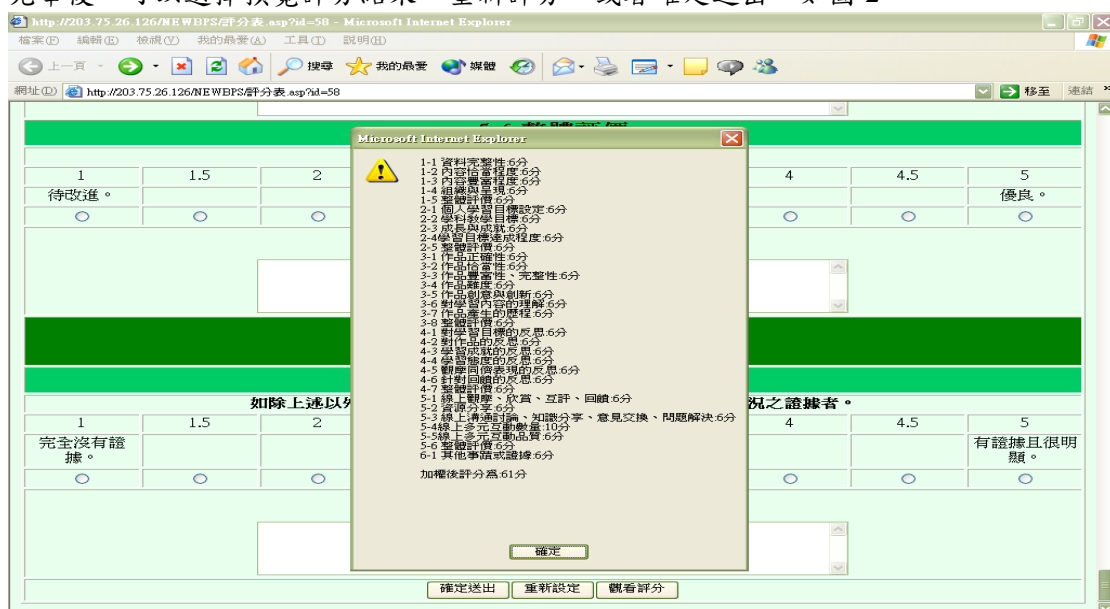


圖 2 評量結果確認

3.7 教學實驗

實驗依課程單元分兩次進行，第一單元為預試，而第二單元為正式施測。如表 4，實驗活動內容主要為學習者課後使用網路化檔案評量系統進行個人檔案的製作（含設定學習目標、撰寫反思、上傳作品等）、觀摩與自我評量，而教學者透過該系統查看學習者的學習狀況與評量學習者的學習表現。上傳的作品包括初期、完成、修正等不同歷程的作品。

為檢測學習者自評效度，學習者課後在兩個單元結束後數日內對自己的學習歷程檔案進行兩次的自我評量（共四次），以確認再測效度。同樣地，教學者（1 位授課教師與 3 位線上助教）亦在兩個單元結束後數日內對每人的檔案做評量。

表 4 實驗活動

週次	對象	活動內容
第一週 (預試階段)	教師	1.在課堂上說明如何製作學習檔案。 2.在課堂上解釋評量標準。 3.準備第一單元課程內容與教學活動。
	助教	1.系統測試。新增學生帳號。 2.在課堂上示範網路化檔案評量系統之使用。
	學生	1.學習如何製作學習檔案。 2.學習如何評分。 3.學習如何使用系統。
第二、三週	教師	進行第一單元課程教學。 上網檢視學生檔案製作狀況。
	助教	進行系統的管理與維護。 上網檢視學生檔案製作狀況。
	學生	課後使用檔案評量系統進行檔案製作。
第四週	教師	進行第一單元課程教學。 上網檢視學生檔案製作狀況。
	助教	上網檢視學生檔案製作狀況。
	學生	使用檔案評量系統進行檔案製作。 上傳初期品與完成作品。
第五週	教師	上網檢視學生作品繳交狀況。
	助教	利用系統發出作品催繳通知。
	學生	上傳修正作品。
第六週	教師	進行教學者評分。
	助教	進行教學者評分。
	學生	進行自評(兩次)、同儕互評。
第七週 (正式施測)	教師	準備第二單元課程與教學活動。 在課堂上說明第一單元實施需要改進的事項。
	助教	整理各項評量數據,並檢驗信效度及修正評量標準。 說明修正的評量標準。
	學生	瞭解第一單元實施需要改進的事項。 瞭解修正的評量標準。
第八、九週	教師、助教、學生	同第二、三週。
第十週	教師、助教、學生	同第四週。
第十一週	教師、助教、學生	同第五週。

第十二週	教師、助教、學生	同第六週。
------	----------	-------

第十三週	助教	整理各項評量數據，並檢驗信效度。
------	----	------------------

4. 結果與討論

4.1 學習者兩次自我評量結果是否一致？

兩次自評結果的 Pearson 積差相關如表 5。結果顯示學習者兩次的檔案自評結果具有高度的相關性且達顯著，表示兩次自評結果具有高度一致性，此結果符合 Zalatan (2001) 指出「學習者對於自己的學習結果評鑑會用心投入，並且發展出穩定的反省思考能力」。各面向中，以檔案製作面向最高，檔案製作面向居次；反思面向最低，顯示反思面向的評審較不容易達成一致。Gadbury-Amyot 等人 (2003) 的研究對紙本式檔案評量效度做檢定，兩位教師評分者之間相關係數介於 0.28 至 0.6 之間，較本研究結果為低。但他們的兩位評分者是教師，本研究是學生自評，評分者的身份不同。

表 5 學習者兩次自我評量結果之 Pearson 積差相關

檔案評量	相關係數 (顯著性)
檔案製作	0.99 (0.003***)
學習目標	0.93 (0.022**)
作品	0.87 (0.011**)
反思	0.78 (0.023**)
態度	0.85 (0.033**)
其它	0.82 (0.028**)
整體	0.81 (0.001***)

註：** $p < 0.01$, *** $p < 0.001$

4.2 學習者自我評量與教學者評量結果是否一致？是否有顯著差異？

學習者自評與教學者評量結果的 Pearson 積差相關及 t 考驗如表 6。結果顯示在作品面向及整體上，兩者具有高度的相關且達顯著，表示該兩種評量方式在作品面向及整體上皆具高度一致性。但在其它的四個面向上，兩者皆未達顯著相關，顯示這兩種評量方式在這四個面向上的一致性低。在假設教學者評量是具有效度而做為效標的前提之下，整體而言，學習者自評的評量方式是具有效度的。

自評與教學者評在檔案製作、作品、態度、整體面向上皆未達顯著差異，顯示兩種評量方法的結果在這四個面向上的落差不大。但在學習目標、反思兩個面向上，自評與教學者評的結果有顯著差異，顯示兩種評量結果在這兩個面向上的落差很大。以差異大小來看（效果量），以反思面向差異最大（效果量），其次為學習目標，檔案製作面向最小。就整體平均分數而言，學習者自評略高於教學者評分，顯示學習者自評較為寬鬆，教學者評分較為嚴格。以各面向平均數而言，除了學習目標、態度之外，其餘三個面向皆為教學者評分高於自評分數。顯示在檔案製作、作品、反思這三個面向上，教學者評分反而較為寬鬆，其原因需要進一步探討。

表 6 學習者自評與教學者評分結果之 Pearson 積差相關及 t 考驗

面向	相關係數 (顯著性)	評分方式	平均數	標準差	t 值 (顯著性)	效果量
檔案製作	0.19 (0.270)	自評	4.02	0.57	-0.309 (0.759)	0.061
		教學者評分	4.05	0.40		
學習目標	0.28 (0.102)	自評	3.93	0.53	2.860 (0.007**)	0.583
		教學者評分	3.69	0.24		

作品	0.49 (0.002**)	自評	3.70	0.44	-0.775 (0.443)	0.138
		教學者評分	3.75	0.26		
反思	0.32 (0.061)	自評	3.82	0.52	-3.289 (0.002**)	0.672
		教學者評分	4.11	0.32		
態度	-0.07 (0.685)	自評	3.62	0.54	1.555 (0.129)	0.379
		教學者評分	3.44	0.40		
其它	0.18 (0.373)	自評	3.54	0.53	0.867 (0.065)	0.105
		教學者評分	3.59	0.42		
整體	0.83 (0.001**)	自評	3.82	0.46	1.058 (0.297)	0.214
		教學者評分	3.74	0.26		

註： 1.** $p < 0.01$ 2.「其它」面向僅有單獨一項，無法計算 Pearson 積差相關係數 3.學習者兩次自評之平均後再與教學者評分的平均做 Pearson 積差相關的檢定

4.3 各項指標是否足以檢測出學習成就(即檔案分數是否與測驗成績一致)? 檔案分數(扣掉作品)是否與作品成績一致?

學習者檔案自互評分數與其測驗成績的 Pearson 積差相關如表 7。結果顯示，學習者的檔案自評分數與其測驗成績具高度一致性且達顯著，個別面向亦具高度一致性且達顯著。顯示檔案自評結果可以反應學習成就，符合 Gelinas (1998) 提列的「檔案評量分數與被評者的學習成果表現有正相關」。其中以作品面向的相關性最高，檔案製作面向居次，其它面向最低且未達顯著。顯示作品最能檢測出學習成就。

表 7 檔案分數與測驗成績之 Pearson 積差相關

面向	相關係數	顯著性
檔案製作	0.67	0.000***
學習目標	0.64	0.003**
作品	0.68	0.000***
反思	0.66	0.000***
態度	0.61	0.006***
其它	0.24	0.008*
整體	0.71	0.023**

註：* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

學習者自評檔案分數(扣掉作品)與作品成績的 Pearson 積差相關如表 8。結果顯示，自評檔案分數(扣掉作品)與作品成績未達顯著相關，顯示兩者一致性低。

表 8 檔案分數與作品成績之 Pearson 積差相關

	檔案分數
作品成績	0.39 (0.275)

4.4 綜合討論

就學者者自評結果的信度而言，表 5 學習者兩次自評結果之 Pearson 積差相關表顯示學習者兩次自評結果具高度一致性。換言之，學習者自評的信度足夠。就自評結果的效度而言，表 6 學習者自評與教學者評結果之 Pearson 積差相關及 t 考驗表顯示學習者自評與教學者評的結果之間具有高度的一致性且無差異。表 7 檔案分數與測驗成績之 Pearson 積差相關表顯示，學習者的檔案自評分數與其測驗成績結果具有高度的一致性。上述兩項結果顯示，學習者自評的效度足夠。綜上所述，學習者自評是一種同時具有信度與效度的檔案評量方式。各面向的信、效度高低順序

並不一致；整體觀之，以作品面向較高，態度與其它面向較低。

表 9 學習者自評相關係數比較表

面向	兩次自評	自評與教學者評	與學習成就相關性
檔案製作	1***	4	2***
學習目標	2***	3	4***
作品	3***	1**	1***
反思	6**	2	3***
態度	4**	6	5**
其它	5**	5	6*

註：1.1 表示最高，依序為 2、3、4、5、6 最低 2.* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

至目前為止，已有一些透過紙本式與網路化檔案進行自評信、效度的研究。儘管這些研究的結果不一，大部份是信、效度足夠的，少部份證明信、效度不足。Sulzen, Young 與 Hannifin (2008) 的研究顯示，數位化檔案評量的信度不足，效度足夠；而增加評分者人數是可以增加評分信度。但無論如何，要讓網路化檔案自我評量具有足夠信、效度是可以做到的。

5. 結論與建議

實驗結果顯示學習者兩次的檔案自評結果具有高度的一致性，因此本研究認為網路化檔案自評是具有足夠信度的評量方式。郭怡雯 (2001) 認為小學生檔案自評是不具信度，但本研究的高職學習者自評是具信度。訓練有素的評分者、對檔案有充分的瞭解可以提升評分的信度

(Gadbury-Amyot, 2003; Oskay, Schallies & Morgil, 2008)。Derham 與 Diperna (2007)、Sulzen, Young 與 Hannifin (2008) 的研究建議，給評分者足夠的訓練可以提升評分的信度。本研究認為年齡所造成的對檔案認知能力差異或心理成熟度可能會對自評信度有所影響。就面向而言，學習者兩次檔案自評結果的一致性以檔案製作面向最高，顯示檔案製作向的評審較容易達成一致；反思面向最低，顯示反思面向的評審較不容易達成一致。

實驗結果顯示學習者自評與教學者評結果具有高度一致性且無顯著差異，因此在教學者評量是有效度的前提之下，學習者自評是有效度的評量方式。就面向而言，只有在作品面向上自評與教學者評具高度一致性；在學習目標、反思面向上，自評與教學者評有顯著差異。具體明確的評量準規是可以提升評分的信度 (Gadbury-Amyot, 2003; Oskay, Schallies & Morgil, 2008)。本研究認為學習者能運用較為具體的評量標準 (作品)，但對於較為抽象評量標準 (學習目標、反思) 的掌握仍須加強。學習者的檔案自評分數與其測驗成績具高度一致性且達顯著，顯示網路化檔案自評可以檢測出學習成就。上述兩項結果顯示，網路化檔案自評是具有足夠效度的評量方式。

由於受限於參與實驗的學習者校內學期進度、畢業考試及升學考量等因素，實驗時間不足三個月，無法進行進行整學期甚至更長時間的實驗研究。由於學習歷程檔案運用於評量上的觀察應以長時間來檢視較佳，因此建議後續的研究可將整個實施時程再予拉長，以增加檔案評量的真實性。同樣地，受限於實驗時間，只採用兩份作品分別進行網路化檔案評量的預試與正式實驗。建議後續研究可以增加作品數量，如此一來每一位學習者可以呈現更多的歷程資料，供評量方式信、效度的檢視。大樣本評量的信度會較高 (Gadbury-Amyot, 2003)。本研究的參與對象屬小樣本，在統計結果的推估上較易有誤差，建議未來研究能增加樣本人數，以提昇研究結果的信、效度。

在反思面向中，考核的重點包含學習者對學習目標、作品、學習成就、學習態度等的反省思考。倘若某位學習者反省內容為「自己是以虛應故事的態度來製作作品，下一次必須更為積極，不可以敷衍了事。」根據評量標準，學習者在反思中明確指出自己的缺點及需改進的地方，所以可獲致較高評價；但是反思內容呈現的是該位學習者學習上的負面缺失，是否應予適當扣分 (本研究顧及學習者擔心扣分而怯於寫出缺點，並未扣分)？或者可另建立哪些指標或機制使評分更加合理，此為後續研究可進一步思考的議題。

參考文獻

王家玲 (2002)。甄選工具之效度驗證與應用-以某高科技公司為例。中央大學人力資源管理研究所碩

士論文，未出版，桃園。

王等元 (2000)。「歷程檔案」概念及其在學習者支持系統之涵義。社會教育學刊，29，249-274。

余民寧 (2003)。教育測驗與評量 - 成就測驗與教學評量 (2 版)。台北：心理出版社股份有限公司。

岳修平、王郁青 (2000)。電子化學習歷程檔案實施之態度研究。教育心理學報，31(2)，65-84。

張麗麗 (2002)。檔案評量信度與效度的分析 - 以國小寫作檔案為例。教育與心理研究，25，1-34。

郭怡雯 (2001)。融入課程的數學檔案評量-以四年級為例。國科會大專生參與專題研究計畫報告，未出版，台北。

陳英豪、吳裕益 (1991)。測驗與評量 (修訂一版)。高雄：復文出版公司。

陳嘉鴻 (2002)。高職資訊性向量表信、效度與結果運用之研究。高雄師範大學資訊教育研究所碩士論文，未出版，高雄。

劉旨峰、林珊如、袁賢銘 (2002)。網路化學習歷程檔案之學習成效分析及未來展望。論文發表於第三屆電子化企業經營管理理論暨實務研討會，彰化，大葉大學。

Chang, C.-C. (2001). A study on the evaluation and effectiveness analysis of web-based learning portfolio (WBLP). *British Journal of Educational Technology*, 32(4), 435-458.

Chang, C.-C. (2008). Enhancing self-perceived effects using Web-based portfolio assessment. *Computers in Human Behavior*, 24(3), 1753-1771.

Chang, C.-C., & Tseng, K.-H. (2009a). Using a Web-Based Portfolio Assessment System to Elevate Project-Based Learning Performances. *Interactive Learning Environments*, 16(2), 25-37.

Chang, C.-C., & Tseng, K.-H. (2009b). Use and performances of web-based portfolio assessment. *British Journal of Educational Technology*, 40(2), 358-370.

Danielson, C., & Abrutyn, L. (1997). *An introduction to using portfolios in the classroom*. Alexandria, VA: Association for Supervision and Curriculum Development.

Derham, C., Diperna, J., (2007). Digital Professional Portfolios of Preservice Teaching: An Initial Study of Score Reliability and Validity. *International Journal of Technology and Teacher Education*, 15(3), 363-381.

Gadbury-Amyot, C. (2003). *Validity and reliability of portfolio assessment of competency in a baccalaureate dental hygiene program*. Unpublished doctoral dissertation, University of Missouri-Kansas City.

Gadbury-Amyot, C. C., Kim, J., Palm R. L., Mills, E., Noble, E & Overman, P. (2003). Validity and reliability of portfolio assessment of competency in a baccalaureate dental hygiene program, *Journal of Dental Education*, 67(9) 991-1002.

Gelinas, A. M. (1998). *Issue of reliability and validity in using portfolio assessment to measure foreign language teacher performance*. Unpublished doctoral dissertation. Ohio State University, Columbus, Ohio.

Hult, C. (2001). Using on-line portfolios to assess english majors at Utah State University. In B. L. Cambridge, S. Kahn, D. P. Tompkins, & K. B. Yancey (Eds.). *Electronic Portfolios - Emerging Practice in Student, Faculty, and Institutional Learning* (pp.60-70), Washington, DC: American Association for Higher Education Press.

Lenze, J. (2004). Inter-rater Reliability in the Evaluation of Electronic Portfolios: A Survey of Empirical Research Results. In R. Ferdig et al. (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference 2004* (pp. 164-169). Chesapeake, VA: AACE.

Lin, S. J., Liu, Z. F., Yuan, S. M. (2001). Web-based peer assessment: Feedback for students with various thinking-styles. *Journal of Computer Assisted Learning*, 17(4), 420-432.

Liu, Z. F., Lin, S. J., Chiu, C. H., & Yuan, S. M. (2001). Web-based peer review: The learner as both adapter and reviewer. *IEEE Transactions on Education*, 44(3), 246-251.

Luca, J., & McMahon, M. (2006). Developing multidisciplinary teams through self-assessment, supported with online tools. In E. Pearson & P. Bohman (Eds.), *Proceedings of Ed-Media* (pp.1855-1860). Norfolk, VA: AACE.

Oskay, O., Schallies, M., & Morgil, I. (2008). A closer look at findings from recent publication. *H. U. Journal of Education*, 35, 263-272.

Russell, J. D., & Butcher, C. (1999). Using portfolios in educational technology courses. *Journal of Technology and Teacher Education*, 7(4), 279-289.

Sluijsmans, D., Dochy, F., & Moerkerke, G. (1999). Creating a learning environment by using self-, peer-, and co-assessment. *Learning Environments Research*, 1(3), 293-319.

Springfield, E. (2001b). Comparing electronic and paper portfolios. In B. L. Cambridge, S. Kahn, D. P. Tompkins, & K. B. Yancey (Eds.). *Electronic Portfolios - Emerging Practice in Student, Faculty, and Institutional Learning* (pp.76-82). Washington, DC: American Association for Higher Education Press.

Sulzen, J., Young, M., & Hannifin, R. (2008). Reliability and validity of an ecologically-grounded student teacher electronic portfolio rubric. In K. McFerrin et al. (Eds.), *Proceedings of Society for Information*

Technology & Teacher Education International Conference 2008 (pp. 153-159). Chesapeake, VA: AACE.

Yancey, K. B. (2001b). General Patterns and the Future. In B. L. Cambridge, S. Kahn, D. P. Tompkins, & K. B. Yancey (Eds.). *Electronic Portfolios - Emerging Practice in Student, Faculty, and Institutional Learning* (pp.83-87), Washington, DC: American Association for Higher Education Press.

Zalatan, J. A. (2001). Electronic portfolios in a management major curriculum. In B. L. Cambridge, S. Kahn, D. P. Tompkins, & K. B. Yancey (Eds.). *Electronic Portfolios - Emerging Practice in Student, Faculty, and Institutional Learning* (pp.37-43), Washington, DC: American Association for Higher Education Press.

新詩賞析平台之研製

Building a modern poetry appreciation platform

楊哲青^a、曾憲雄^b、王碧玲^a、楊智凱^b、翁瑞鋒^a、蘇俊銘^a

^a國立交通大學

^b亞洲大學

jerome@cis.nctu.edu.tw, sstseng@asia.edu.tw, isfion@hotmail.com, moto41@asia.edu.tw,
roy@cis.nctu.edu.tw, jmsu@csie.nctu.edu.tw

【摘要】 新詩教學在指導學生捕捉詩中意象，也就是瞭解新詩中的形式以及情感美與心靈美。但由於新詩長短不一，運用多種結構及修辭技巧，導致較難學習。為使學生能學習賞析新詩的形式及內容賞析，我們提出並建構賞析本體論，並透過設計「範文引導形式賞析」及「意圖標籤內容賞析」之賞析標籤為主的學習輔助系統，評量學生的學習。本實驗施測於高中學生與國文教師，實驗發現藉由賞析標籤，可以讓學生了解新詩的章法形式結構，誘發其對於新詩意象的想像空間及感覺。學習不再受限於教師教學與課堂時間，亦提昇新詩整體學習成效及滿意度。

【關鍵詞】 新詩賞析、本體論、論文引導形式賞析、意圖標籤內容賞析

Abstract: Modern poetry teaching focuses on how to instruct students to catch the imagery in the poem, which includes form understanding and content appreciating. Through learning, students can realize the beauty of emotion and spirit. However, modern poetry usually has complex structures with various lengths and rhetoric skills, and appreciating the contents of modern poetry needs more background knowledge. Therefore, students are usually unable to realize the meaning and imagery of modern poetry. In order to motivate students to learn understanding and content appreciating of modern poetry, we analyzed the methods of appreciation to build the appreciation ontology. In addition, the model essay teaching was conducted to guide the student to appreciate the form of modern poetry. We proposed “model essay guided form appreciation” and “intention finding for content appreciation” to assist learner appreciating modern poetry. The experiment conducted the questionnaire analysis for students and teacher interview. To evaluate the performance of this system, students are invited to do questionnaire survey and Chinese teachers of a high school are invited to do teacher interview. Through this system, students not only understand the form of modern poetry, but also bring out the imagination and feeling of the modern poetry. In addition, learning can be conducted from anywhere and anytime. The interactions of teachers and students can further assist teachers assess the learning achievement of students easily.

Keywords: modern poetry appreciation, ontology, model essay guided form appreciation, intention finding for content appreciation

1. 簡介

在浩瀚的文學領域中，詩具有獨特的智慧及藝術形式。詩的美在於它給人一種無限想像的空間及一份不可言喻的韻味，讓人可以長久的低迴品味而歷久彌新。在語文教學當中，詩的教學常是脫離不了讀、說、寫、作這四個步驟，也就是從閱讀來獲得知識，再走向欣賞及創作之路，再經由詩的創作來提昇寫作能力與技巧，讓豐富的情感透過詩的創作，來

盡情展現。因此，懂得閱讀及賞析新詩，進而從中模仿與創作學習，對於提升語文表達能力多所助益。新詩賞析，除基礎知識教學外，「意象」更為重要。「意」指內在的心意，「象」為外在的景象。內在之意需借外在具體事物、行為、感官等「象」來表達，「意象」是將內在的情感具體化的過程，然而，為使意象呈現多種面貌，其轉化是需透過各種修辭技巧來展現的。讀詩、解詩，首重文本的詮釋鑑賞，欲掘發詩人的情感與思想，洞澈詩境，可由詩的章法結構與修辭技巧切入(曾進豐，民 96)。因此，詩在創作方法上為了讓文字不露骨，不流於直說的率直的表達方式，於是運用各種修辭技巧，讓詩意與詩句多所轉折並呈現出多樣的面貌(李翠瑛，民 98)。

本研究利用新詩賞析本體論來設計出新詩賞析平台，透過範文引導教學、斷詞系統(中央研究院中文斷詞系統，民 99)與同義詞詞林(梅家駒，民 86)等等的方式，以資訊技術結合中國文學詩作，提供一個新的方法來教授與體會新詩。學生可依此標準進行新詩「形式」及「內容」賞析，提供學生自主學習；教師可將同學的賞析詩作之成果共享，提供其他教師做為教學參考，進而協助教師在新詩教學時能有所突破，並提升教學效果。

2. 研究背景

新體詩常見的類型有新詩、白話詩、自由詩及現代詩等。新詩，相對於「舊詩」，打破舊詩格律、形式的限制，句數、字數不限，平仄不拘，押韻與否也無嚴格章則。在詩的內涵與視境上、表現技巧與語言魅力方面，五、六十年來，詩人們不斷地實驗與努力，在詩田中默默耕耘，貢獻其心血，使「新詩」在各種文類中出類拔萃，與古典詩爭輝。

新詩的教學是情性也是美感的教學，除了知識的教學及習慣技能之教學外，應更重視「鑑賞」，了解作者當時的創作背景、心境、意圖等就可循著文字的歌詠，依內容、主旨、情意、作法、修辭等加以一一鑑賞，一窺文學作品的不同風貌，從文字的內涵和外在形式風格得到新詩的意蘊美。而新詩鑑賞教學的重要性，是要實現文學鑑賞活動的三大社會功能，可達到潛移默化、寓教於樂及以情感人的美育作用(孔佳薇，民 97)。因此，我們使用範文來引導學習，實現新詩鑑賞教學目的，範文是學習模範的好文章，可作為學生學習時的範本(黃錦鉉，民 86)。

隨著 Web2.0 網路科技發展，讓使用者透過標籤(tag)來輔助數位文件的分類整理(Wang, 2010)，因為使用者可以清楚的表達出對文件的觀感，因此近年來被廣為使用。而本體論則是常被用來描述單一領域標籤背後的語意關連，常見的本體論建造方法如 Mike Uschold (1996)等的骨架法、Michael Gruninger (1995, 1998)等的 TOVE 評價法及 Noy, N. F. (2001)等提出的本體工程方法。然而若要用在描述新詩賞析上，則需參考新詩賞析策略(陳啟佑，民 90a，民 90b)，並針對新詩常用賞析字彙建置描述標籤，因此如何建置具有好的描述力，並能提供診斷效果的賞析本體論，是本論文重要的技術議題。

在擴增賞析本體論部份，《同義詞詞林》是由上海外語學院梅家駒等人共同編輯的中文詞義分類工具書(梅家駒，民 86)，總共分 12 大類，94 中類，1428 小類。在小類之下又根據同義的原則劃分不同的詞群，共計有 3,925 個詞群，包含將近七萬個詞義項目。由於《同義詞詞林》具有同義詞彙可供參考，使用其協助學習者釐清想法標示新詩。

3. 新詩賞析研究方法

本研究主要建置新詩賞析學習平台來讓學生進行新詩賞析活動，新詩雖不拘格律，事實上，仍應有其基本賞析準則架構。因此，為了能讓學生透過標籤來描述出對新詩的賞析結

果，本研究參考相關新詩賞析文獻(李翠瑛，民 98；林文欽，民 89；陳啟佑，民 90a，民 90b)，建構賞析本體論，再藉由賞析本體論去評量學生在新詩賞析學習活動中的賞析結果。

3.1. 研究目的與方法

新詩賞析包含了新詩形式賞析與新詩內容賞析，並以「問題－答案」模式進行。在形式賞析部分，學習者可能觀念錯誤或概念混淆，因此，我們利用國語文教學上常使用的範文教學，來引導學習者找出形式賞析，在此我們稱這為「範文引導形式賞析」。在內容賞析部份，為取得學習者賞析的意圖，協助教師了解學生看法與感受，本研究利用斷詞系統與同義詞詞林協助學習者更精確的描述賞析結果。稱為「意圖標籤內容賞析」。最後，學習者賞析結果一方面提供其他學習者學習參考，另一方面給予教師進行教學評量，進而瞭解學生學習狀況，以便調整教學方法。新詩賞析標籤尋找流程如圖 1 所示。

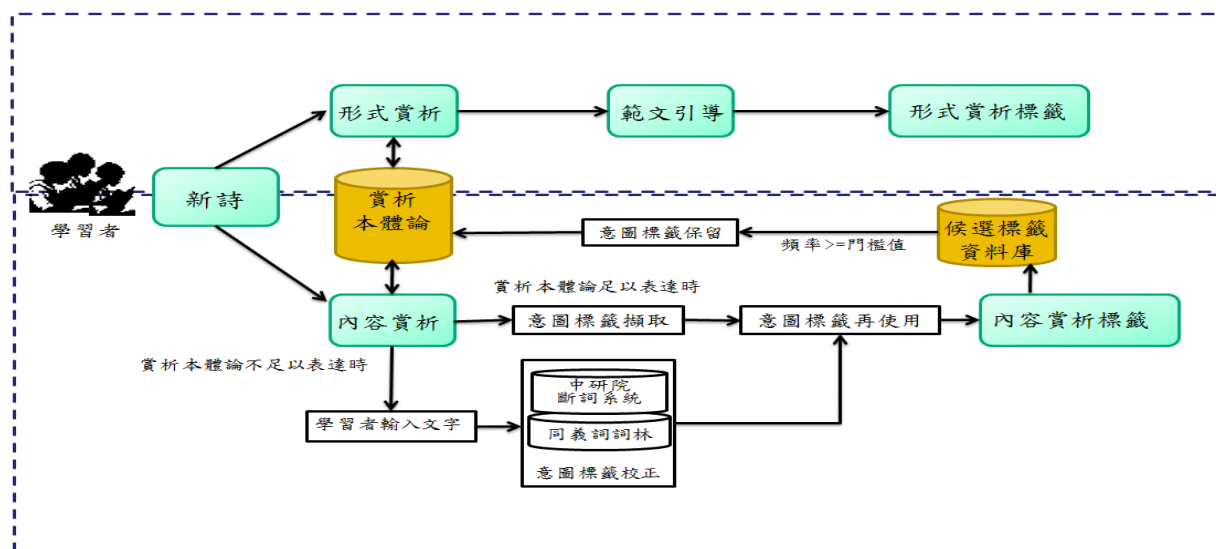


圖 1 新詩賞析標籤尋找流程

3.2. 賞析本體論之建置

本研究為輔助學生進行新詩賞析之評論，因此建置賞析本體論。此本體論參酌各學者所提建構知識本體論之步驟及新詩賞析相關文獻而建置，分為以下四個步驟進行：

(1). 決定研究領域與範圍

利用「階層詢答集」的方式來進行，高層級為問題，回答之答案則為低層級。本研究依此「問題－答案」模式進行本體之建置，例如「本首詩作中所要進行賞析為何？」、「本首詩作中是否有下列修辭技巧？」、「本首詩作中形式結構章法為何？」，藉由問題與答案的連結逐步形成本體架構。

(2). 列舉知識本體中的重要詞彙

參考相關學者所提新詩賞析策略，綜合歸納分為形式賞析及內容賞析二部分。

(3). 定義類別及階層

將蒐集的詞彙由上而下進行分類整理，由範圍最廣之「新詩賞析」領域開始，其下接子類別「形式賞析」「內容賞析」等概念，「形式賞析」之下又可分「結構形式」「修辭技巧」「聲律」，為「形式賞析」之子類別，以此類推逐步建立類別及階層。

(4). 建立及表達本體知識

賞析本體論架構如圖 2 所示，在新詩賞析學習活動過程中，透過該賞析本體論，可以用來輔助學生標示對新詩的賞析模式。

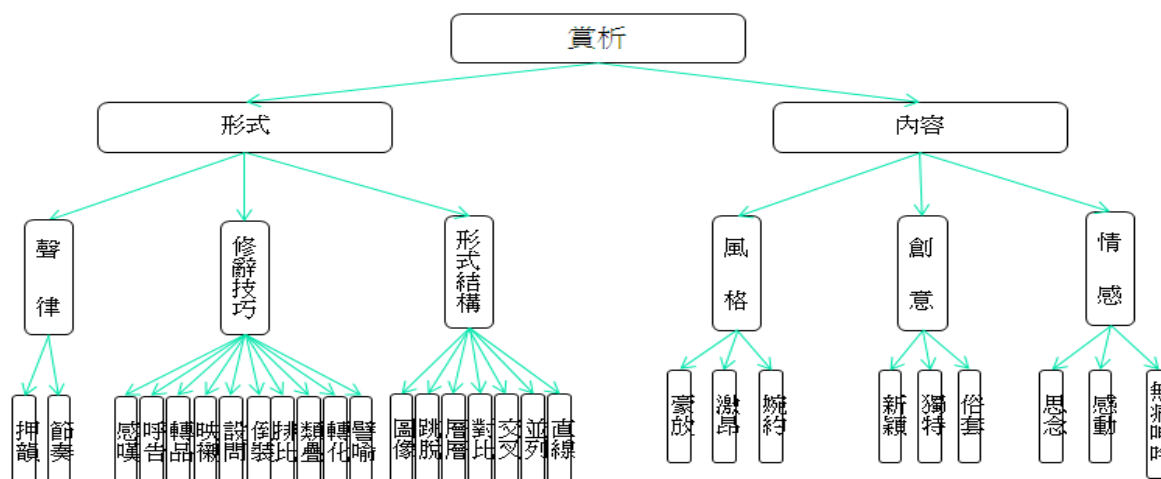


圖 2 賞析本體論架構圖

3.3. 新詩形式賞析

針對形式賞析，我們利用建立的賞析本體論根據定義的根節點所對應的子節點，以詢答集「問題—答案」方式進行，並於詢答過程中透過範文引導，來刺激學生思考使之更容易掌握、觀察出一首詩作所存在的形式賞析技巧。形式賞析可分整體架構與部分詞句來進行賞析，以整體架構賞析而言，例如：新詩的形式結構賞析，由學習者觀察後認為形式賞析為直線法。以部分詞句而言，例如：修辭技巧中的倒裝，學習者觀察後認為形式賞析為倒裝句法，並指出詩作中的哪些詞句為倒裝句。以下圖 3 為新詩形式賞析活動範例流程。

範例 1. 新詩形式賞析活動範例

鄭愁予

錯誤

那等在季節裡的容顏如蓮花的開落

東風不來，三月的柳絮不飛

你的心如小小的寂寞的城

恰若青石的街道向晚

跫音不響，三月的春帷不揭

你的心是小小的窗扉緊掩

我達達的馬蹄聲是美麗的錯誤

我不是歸人，是個過客

透過演算法設計，新詩範文引導形式賞析學習活動為：

系統：「形式」部分，你想描述以下哪一部分呢？

形式結構、修辭技巧、聲律

學生：選擇所要賞析的項目為：修辭技巧

系統：就「修辭技巧」而言，您認為該詩作運用了哪些技巧？

譬喻、轉化、類疊、排比、倒裝、設問、映襯、轉品、呼告、感嘆

學生：選擇所要賞析的項目為：倒裝法

如此進行詢答後，則系統會有對於倒裝法的說明與範文引導如：

解釋：倒裝句為特意顛倒語句文法上、邏輯上正常順序的修辭法。

範文：「所以，我去，總穿一襲藍衫子」〈情婦〉

解析：「所以，我總穿一襲藍衫子去」

圖 3 新詩形式賞析活動範例

透過引導說明，協助學習者能觀察出詩作中的倒裝句法且將在位於文章內容何處並且標示出來，以「錯誤」這篇為例，學生標示出來的倒裝句為：「你的心是小小的窗扉緊掩」，最後系統將記錄學習者的賞析內容，其記錄為：形式 → 修辭技巧 → 倒裝法（你的心是小小的窗扉緊掩）。

3.4. 新詩內容賞析

針對內容賞析方面，透過賞析本體論找出學習者對於新詩的情感、創意、風格等看法。學習者根據賞析本體論所存在的重要詞彙進行賞析，例如就風格而言：重要詞彙有婉約、激昂、豪放等，當賞析本體所定義的重要詞彙不足以表達學習者對於詩作的感受時，學習者須自行輸入一段對於詩作看法的描述文字，系統將會透過中研院斷詞系統解析出詞性，本研究擷取名詞、動詞、形容詞、副詞做為學習者意圖表達的關鍵字，在經由學習者選擇意圖表達的關鍵字，透過同義詞詞林協助學習者找出更精確詞彙來表達感受或看法。意圖標籤內容賞析運作如下：

- (1).意圖標籤擷取：當賞析本體論所定義的重要詞彙足以表達學習者對詩作整體的風格、題材、描寫手法、情感，學習者透過選項選擇出所要表達的詞彙即可。
- (2).意圖標籤校正：當賞析本體論所定義的重要詞彙不足以表達學習者的感受時，如：很好、不錯等，此種無法具體表達對於詩作的感受或想法，所以我們需要學習者輸入對詩作的看法，藉由輸入之文字，進行斷詞動作，並配合同義詞詞林推薦詞彙，協助學生能更明確的表達出對於詩作賞析後的感受。
- (3).意圖標籤再使用：學習者透過意圖標籤擷取及意圖標籤校正的方法，所取出的詞彙，來表達對於詩作的情感、創意、風格等看法。
- (4).意圖標籤保留：學習者透過文字輸入，經斷詞系統及同義詞詞林明確表達對詩作賞析的感受後，系統會將之詞彙存於候選標籤資料庫，希望透過眾人之力不斷的累積賞析本體論知識，其中包括對於詩作的情感、創意、風格等看法，為了避免賞析本體論知識無意義的累積，候選標籤表達的詞彙需超過某一門檻值，才能加入賞析本體論中。

以下圖 4 同樣以鄭愁予的新詩「錯誤」來進行新詩內容賞析活動範例流程。

範例 2. 新詩內容賞析活動範例

系統：「內容」部分，你想描述以下哪一部分呢？

情感、創意、風格

學生：選擇所要賞析的項目為：情感

系統：就「情感」而言，您認為此首詩作表達情感意境為何？

無病呻吟、感動、思念、以上皆非

學生：選擇所要的項目若為：感動

則系統將會記錄學習者的賞析內容為：內容 → 情感 → 感動

學生：選擇所要的項目若為：以上皆非

則系統將會要求學習者輸入自然語言，如學習者輸入「從期待到最後落空」經由斷詞系統後過濾贅字，找出關鍵候選字如「期待、落空」，學習者在選擇所要表達的關鍵字如「落空」，在經由同義詞詞林找出「落空」的大、中、小類概念如：一場春夢、一場空、功敗垂成、南柯夢等，學習者在從此這幾個詞彙選擇所要表達的感受如：南柯夢，最後系統將會記錄學習者的賞析內容為：內容 → 情感 → 南柯夢。

圖 4 新詩內容賞析活動範例

4. 系統實作

本研究依據新詩賞析本體論為基礎建置了新詩賞析系統平台，學生可依此標準進行新詩的形式及內容賞析，教師則可將賞析成果作為評量或教學回饋工具。除此之外，學習者亦可線上觀看教師評量後的賞析結果及其他同學的評析結果，共享學習之成果及砥礪反省之用。

4.1. 系統功能介紹

形式賞析：當開始進行活動前，學習者必須輸入自己的姓名或代號，以登入詩作賞析畫面。登入後，學習者可圈選欲先賞析的面向，爾後逐步展開賞析活動，若對於賞析面向有不清楚之處，請將滑鼠移至【說明】處，即會顯示該項目之範文及解釋。

所謂形式賞析，包括詩作的結構、斷句、分行、修辭、文法、押韻等。形式賞析其下分為三部分，分別為：「形式結構」、「修辭技巧」及「聲律」。下圖為形式結構範例，形式結構為詩作展開的整體，分為「直線」、「並列」、「交叉」、「對比」、「層層」、「跳脫」及「圖像」等七部份，可供學生選擇來標示對於詩詞文句之賞析，如圖 5 所示。

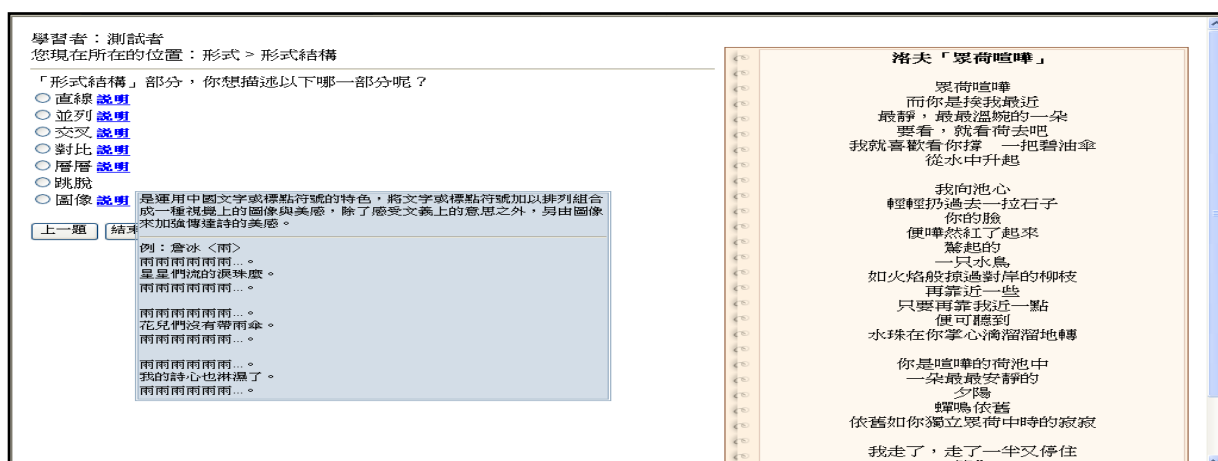


圖 5 學生新詩賞析功能－形式結構賞析

4.2. 實驗設計

本研究的實驗程序包含四個步驟：(1)進行傳統紙本新詩賞析活動、(2)簡介新詩賞析系統內容及流程、(3)進行線上新詩賞析活動、(4)填寫問卷。實驗對象為臺北市某私立高中二年級之學生，採便利樣本，受試者共三班共計有 105 人。其中男生 51 位女生 54 位，年齡分布 16 歲(含)以下 13 位、17 歲 85 位、18 歲 7 位，老師部分則是同學校國文科專業教師共 3 人。

本研究將新詩賞析學習滿意度問卷類型分為形式賞析、內容賞析及系統可用性等三個構面，共計 15 題。係採李克特 (Likert) 五點量表，選項依序為「完全同意」、「有點同意」、「沒意見」、「有點不同意」及「完全不同意」五個選項。讓受試者依自己的情形作答，從選項中勾選一個最符合的答案，如表 1。

表 1：新詩學習滿意度問卷構面

衡量構面	題項	問卷題目
	1	對於範文引導解說，我可以清楚明白並瞭解其所要表達的意義。

形式 賞 析	2	運用範文引導進行新詩賞析的方式，能協助我觀察對照並指出詩作的形式結構。
	3	運用範文引導進行新詩賞析的方式，能協助我觀察對照並指出詩作的修辭技巧。
	4	運用範文引導進行新詩賞析的方式，能協助我觀察對照並指出詩作的聲律。
	5	透過系統的引導說明，我可以知道如何進行新詩賞析活動。
	6	運用範文引導進行的方式有助於我學習新詩賞析。
內 容 賞 析	7	透過新詩內容賞析，能讓我能瞭解其他同學對於新詩賞析的不同看法。
	8	透過斷詞系統及同義詞詞林可以協助我找出所要表達的感受。
	9	透過同義詞詞林可以讓我學習到更多的相關用語及詞彙。
	10	透過新詩內容賞析，讓我學習到可從情感、創意、風格這些面向進行賞析。
系 統 可 用 性	11	我覺得新詩賞析平台介面操作相當容易。
	12	我覺得透過新詩賞析平台來進行學習活動是容易的。
	13	我希望能繼續透過新詩賞析平台的方式來進行新詩賞析的學習。
	14	整體而言，在平台上進行的學習方式可以輔助我對於新詩賞析的學習。
	15	整體而言，我對這次在平台上進行的新詩賞析學習方式感覺滿意。

5. 問卷資料分析

在項目分析與信度分析部分，檢驗新詩學習滿意度量表與總分的相關情形，十五題題項與總分的相關均為正相關，且均達顯著水準。全量表的內部一致性信度 Cronbach α 值為.908。研究變項敘述統計分析，根據敘述統計分析發現，整體來說平均數均高於3，最高4.01 最低3.70，整體平均為3.814，標準差為0.91，同意程度偏高，代表受試者對於系統的滿意度有正向態度。透過學生滿意度問卷的調查，顯示所建構的新詩賞析系統，從情感、創意、風格這些面向進行賞析的滿意度最高，而文章結構、修辭聲律之形式賞析構面次之，其餘學習者對於範文引導活動對賞析的幫助與系統可用性構面等，其滿意度平均數均高於3，表示受測的105位學生，對於整體系統功能及協助學習上滿意度同意程度偏高，具正向態度。

此外本研究共訪問3位國文教師，針對新詩賞析平台的使用觀點展開對談。受訪者皆指出學生對於新詩賞析平台，有別於傳統枯燥的學習方式，能協助閱讀能力不足的學生，透過更多詞彙來練習表達能力。以下結錄部分訪談教師對系統的教學心得：

- 教師 A:『學生對於感覺的述說並不擅長，加上現在一些特殊流行語的盛行，學生的閱讀及寫作能力更差，因此加上斷詞系統以為輔助，學生可以知道有更多詞句可用，這樣可協助學生學習更多詞彙，亦可讓學生了解自己的思考方向太過狹隘。』
- 教師 B:『能夠提供辭句的練習，讓老師掌握學生學習的狀況，藉此瞭解學生不足或需改進的地方，尤其透過系統配合同義詞詞林找出詞彙，可以豐富學生的用語。』

總結實驗結果，新詩賞析平台對於學生是有幫助的，透過引導可以讓學生明白其意義協助賞析，且該系統能提供學生自我檢測，也讓老師可以掌握學生的程度，以做為教學參考指標。此外教師並建議可以搭配多媒體素材如：小遊戲、音樂等，來給予鼓勵並加深情境學習。並可以開放學生交流，來更豐富學習的活動內涵。

6. 結論

本研究建構新詩賞析本體論概念，應用此概念設計新詩賞析平台系統，利用範文引導教學、斷詞系統及同義詞詞林，即是期望透過資訊結合新詩賞析教學，藉由網路讓學生學習如何賞析，表達對於詩作的看法，也可以讓教師瞭解學生的學習狀況與學生對於作品的想法，透過彼此的互動教學相長。經由實驗證明，藉由範文引導說明了解新詩形式結構、修辭技巧與聲律對於學習者是有良好的學習成效，而使用斷詞系統及同義詞詞林擷取出的詞彙，也能誘發對於新詩意象的想像空間及感覺描述。研究的賞析活動過程中，解決師生互動不頻繁的問題，教師可以任意指派新詩以為課堂或課後閱讀補充教材，並反覆要求學生進行練習，對於賞析後的結果，教師也能適時回饋並且評量，並從賞析活動中瞭解個別學生之學習問題，輔以專屬或加強教學，讓教學兼顧個別差異進而因材施教。

本研究針對新詩賞析有了初步成果，未來系統亦可延伸多媒體教學設計，讓新詩的賞析活動更活潑且多元。此外，系統可增加更多的互動性，如闖關遊戲的設計、同儕互評機制等，提高學生學習興趣與樂。懂得鑑賞後，更期望能進一步鼓勵新詩創意的寫作，致使語文教學中，基本能力「聽、說、讀、寫」四步驟都能均衡發展。

致謝

本論文承蒙國科會計畫部分補助，計畫編號 NSC97-2511-S-468-004-MY3、NSC98-2511-S-468-004-MY3、NSC97-2511-S-009-001-MY3、NSC95-2520-S009-008-MY3。

參考文獻

- 中央研究院中文斷詞系統（無日期）。擷取自民 99 年 4 月 16 日，於
<http://ckipsvr.iis.sinica.edu.tw/>
- 孔佳薇（民 97）。新詩教學的探究—以現行高中國文教材為例。國立臺灣師範大學國文研究所碩士論文，未出版，台北市。
- 李翠瑛（無日期）。現代詩中的意象經營。民 99 年 4 月 16 日，擷取自
<http://www.docin.com/p-4190256.html>
- 林文欽（民 93）。現代詩教學中章法形式深究探研。高雄師大國文學報，1。
- 梅家駒（民 86）。同義詞詞林。台北市：台灣東華書局股份有限公司。
- 陳啟佑（渡也）（民 90a 年 6 月）。新詩賞析策略(上)。國文天地，17 卷 1 期，64-70。
- 陳啟佑（渡也）（民 90b 年 7 月）。新詩賞析策略(下)。國文天地，17 卷 2 期，50-56。
- 曾進豐（民 96）。淺談新詩教學的理念與策略。國文新天地，15 期，6-11。
- 黃錦鉉（民 86）。國文教學法。台北市：三民書局。
- Fox, M.S., & Gruninger, M. (1998). Enterprise modeling. *AI Magazine*, 19(3), 109-121.

- Gruninger, M., Fox, M.S. (1995). *Methodology for the Design and Evaluation of Ontologies*. Paper presented at the Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95, Montreal.
- Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology (Tech. Rep. KSL-01-05). Stanford University, Knowledge Systems, AI Laboratory.
- Uschold, M., & Gruninger, M. (1996). ONTOLOGIES: Principles, Methods and Applications, *Knowledge Engineering Review*, 11(2), 122-147.
- Wang, J., Clements, M., Yang, J., de Vries, A. P., & Reinders, M.J.T. (2010). Personalization of tagging systems. *Information Processing and Management*, 46(1), 58-70.

A Review of the Strategies for Output Correctness Determination in Automated Assessment of Student Programs^{*}

Chung Man Tang, Yuen Tak Yu[†], Chung Keung Poon

Department of Computer Science, City University of Hong Kong

Email: cmtang.cs@student.cityu.edu.hk, csytyu@cityu.edu.hk, ckpoon@cs.cityu.edu.hk

Abstract: Automated program assessment systems are very useful in enhancing the learning of computer programming, but they typically suffer from technical limitations in the determination of the correctness of student program outputs. This paper reviews the various strategies used in practice to address the problem, contributes to better-informed evaluation of different solutions, and highlights recent advances that are promising in improving existing systems and reducing the efforts spent by instructors.

Keywords: Automated assessment system, pattern-oriented software testing, program assessment requirements, program testing and validation, token pattern

1. Introduction

The worldwide trend of large classes in computer programming courses has stimulated the development of automated learning and assessment systems in many universities (Ala-Mutka, 2005). These systems vary in their capabilities, but they are popularly used for automatically assessing student programs (Higgins et al., 2002; Joy et al., 2005; Nazir et al., 2005). In practice, since holistic assessment of student programs cannot be fully automated, existing systems typically avoid those aspects of programs that are hard to be assessed automatically and objectively, such as programming style and the interpretation of comments in code (Yu et al., 2006).

One common aspect assessed by automated systems is the functional correctness (or simply *correctness*) of students' programs, typically by means of testing. Automation of test execution is relatively unproblematic; however, determination of the correctness of program outputs is far from straightforward (Luck & Joy, 1999). The primary difficulty is that different correct (or acceptable) solutions to the same programming exercise may not always produce exactly the same output (Tang et al., 2009a). Thus, a student program manually judged by the instructor to be correct might be inappropriately rejected by an automated assessment system. This technical limitation definitely needs to be addressed as it has given rise to many educationally undesirable effects in teaching and learning that can substantially compromise these systems' benefits (Tang et al., 2009b).

While the output correctness determination problem has been widely recognized, existing work tends to ignore or avoid it, or consider it lightly often as a side issue in conjunction with a bunch of other unrelated issues and with little reference to others' solutions. In the literature, reports on strategies to deal with the problem have been scattered, rendering them hard to be evaluated by educators and researchers. This paper serves to provide a concise review of the varied strategies in use, and highlights recent advances in this regard.

^{*} This work is supported in part by grants (project numbers 123206 and 123207) from the Research Grants Council of the HKSAR, China.

[†] Corresponding author.

2. The Problem and an Example

In assessing the correctness of a student program automatically by means of testing, the most critical problem is to determine if the program's *actual output* in every test run is correct. The output correctness determination problem is known as the *test oracle problem* in software testing (Zin & Foxley, 1996; Shukla et al., 2005).

Without automation, the instructor tends to either browse through the program listing looking for obvious faults and, if at all, execute each program manually against only a few test cases. Thus, output correctness is judged manually by the instructor, and he/she ultimately decides the correctness criteria. Manual assessment is not only tedious, but also error-prone. Moreover, the correctness criteria often become inconsistent, particularly when the workload of assessing a large number of programs is shared among different assessors.

On the other hand, automated assessment systems will be tireless, able to assess a large number of programs using a much larger set of test cases and with strictly consistent criteria (Nazir et al., 2005). However, currently most systems employ the *output comparison method*, which works by matching the actual output texts with the expected (Ala-Mutka, 2005). Unfortunately, a naïve implementation of the method would render the assessment too inflexible, since any slight deviations from the expected output would not be tolerated (Luck & Joy, 1999). Even when enhanced with simple filters or sophisticated parsing tools, many existing systems are still complained by students to be “*too fussy*” or “*too picky with spaces*” (Joy et al., 2005), or to cause frustration and confusion, as evidenced by comments like “*Sometimes it is right to you but wrong to the automark*” (Suleman, 2008).

To illustrate some common issues in output correctness determination, we shall use the programming exercise (Exercise 1) in Figure 1 as a running example throughout this paper.

When a student's program is executed with the sample input in Figure 1, the actual output is of course correct if it is exactly the same as the sample output, but it is not unusual for another correct program solution to give a slightly different output, such as those produced by programs Prog-A, Prog-B and Prog-C in Figure 2.

Notice that all three programs in Figure 2 compute the correct answers but produce slightly different outputs: (1) Prog-A outputs all words in lower case and omits many blank spaces, (2) Prog-B prints a colon instead of the equal sign and the full name of units instead of their abbreviations, and (3) Prog-C misspells the word *Average* and expresses the number of seconds in one decimal place. Although the output of Prog-A is not very legible due to the missing blanks, and the misspelling of the word *Average* by Prog-C is clearly a mistake, most instructors would consider that these deviations are relatively insignificant and the three programs are acceptable. A human assessor can therefore exercise discretion to mark these programs as correct (perhaps imposing a small penalty). An automated assessment system, however, is not always able to recognize these deviations as insignificant and, consequently, may simply reject these programs as incorrect, resulting in a much heavier penalty. This mismatch between human and automated assessment frequently causes students' frustration (Tang et al., 2009b).

Exercise 1 Write a program that reads a distance (in kilometers), followed by the average speed (in kilometers per hour) travelled in a journey, and calculates the time (in hours, minutes and seconds) taken for the travel.

Sample input
125 90

Sample output
Distance = 125 km
Average speed = 90 km/h
Time taken = 1 h 23 min 20 s

Figure 1. An example programming exercise.

Prog-A	Prog-B	Prog-C
distance=125km average speed=90km/h time taken=1h23min20s	Distance: 125 kilometers Average speed: 90 kilometers/hour Time taken: 1 hours 23 minutes 20 seconds	Distance = 125 km Averge speed = 90 km/h Time taken = 1 h 23 min 20.0 s

Figure 2. Actual outputs of some programs corresponding to the sample input “125 90”.

3. Strategies for Output Correctness Determination

3.1. Basic Character Matching

A primitive implementation of output comparison is to match the actual and expected outputs character by character, using system utilities such as `diff` (Luck & Joy, 1999) or `cmp` (Harris et al., 2004). Thus, basic character matching considers an actual output correct if and only if it is exactly the same text string as the expected output. This method works adequately for programs that, by nature, demand strict conformance to exactly one correct output. Examples are encoding of input texts into a compressed string using a given deterministic algorithm, or into bit streams that satisfy a certain protocol for communication through a network.

3.2. Simple Character Filtering and Conversion

Most exercises in programming classes tend to mimic daily applications where many variants of outputs are acceptable. In particular, minor deviations (such as different character cases or spacing) are commonly tolerated. Thus, in practice, basic character matching is seldom used alone, but is usually supplemented by some simple character filtering and conversion rules. For example, **TRY** includes a utility program `try_deblank` (Reek, 1989) to make the output comparison less sensitive to blanks and empty lines, and **BOSS** (Luck & Joy, 1999) uses a Unix Shell script to preprocess program outputs to ignore whitespace and case sensitivity. This approach is still commonly used in many existing systems. However, the rules for such processing are ad hoc and unsatisfactory, and different exercises may require different rules. **PASS** (Yu et al., 2006), for instance, provides a list of rules from which the instructor may select, such as filtering only the beginning empty lines.

3.3. Prescribing Highly Detailed Specifications

Many instructors spend extra efforts to avoid the output variation problem by prescribing unusually detailed and highly precise output requirements, hopefully to ensure that the correct outputs are unique. Students are also warned in advance that their programs will be rejected by the system unless strict conformance is achieved. For example, a student documentation for the Curator Grader (Curator, 2009) explicitly warns, “If you do have extra lines or missing lines, then the Grader may compare the wrong lines and you will receive a very low score.” It even stresses, in bold type font, the need to “**follow the project specifications for output precisely!**”

Specifying the outputs very precisely and demanding strict compliance to formatting requirements may reduce, but not eliminate, misunderstanding. Moreover, an overly detailed specification can be time-consuming to define, and can become so restrictive that it inhibits creativity, distracts students from the essentials of the exercise, and sometimes is infeasible for certain types of exercises (Jackson, 1991). The “warned-you-before” strategy helps reduce students’ complaints, but not their frustration (Tang et al., 2009b).

3.4. Instructor-provided Stub

Other than preprocessing, one way to achieve uniform output format for easy match by an automated system is to provide the interface for a stub (an instructor-defined function) that students must invoke when producing output from their programs (Reek, 1989). Figure 3 shows a sample stub for Exercise 1. Figure 4 shows an extract of a student program that invokes the stub. In a sense, this strategy not only defines a highly precise output format (in the form of code), but actually obviates the need for students to code the output.

```
void Display::printOutput(int distance, int speed, int hour, int minute, double second){
    cout << "Distance = " << distance << " km" << endl;
    cout << "Speed = " << speed << " km/h" << endl;
    cout << "Time taken = " << hour << " h ";
    if (minute < 10) cout << "0"; // output a '0' before single digit minute output
    cout << minute << " min ";
    if (second < 10) cout << "0"; // output a '0' before single digit second output
    cout.setf(ios_base::fixed, ios_base::floatfield);
    cout.precision(2); cout << second << " s" << endl;
}
```

Figure 3. A stub program to be invoked by students to produce output for Exercise 1.

```
int main(){
    int distance, speed, hour, minute; double second;
    Display disp; // stub provided by instructor
    ... // student's code to compute the time taken in hour, minute & seconds
    disp.printOutput(distance, speed, hour, minute, second); // invokes instructor's stub
}
```

Figure 4. A student program that invokes the stub in Figure 3 to produce output.

3.5. Instructor-designed Driver

The instructor can make up exercises that require students to code a function or partial program instead of a complete program that produces console output (Ala-Mutka, 2005). Thus, output checking reduces to verifying the return values of well-defined data types from the function. The instructor has to design a driver that invokes students' code and then either (1) directly verifies students' computed return values, or (2) outputs the return values in a uniform format for automatic output comparison. The driver serves as a test oracle in the former case and a *wrapper* in the latter case (Shukla et al., 2005). Systems that adopt this strategy include HoGG (Morris, 2003) and Scheme-robo (Saikkonen et al., 2001). The wrapper/stub strategies can also be adapted to assessment of graphical user interface (GUI) programs by converting their I/Os into text streams (English, 2004).

While the driver/stub strategies are commonly used in data structure or object-oriented programming courses which stress the separation of presentation and computation logic (Tremblay et al., 2008), they are not applicable to all exercises. For example, for exercises like "write a program to generate a multiplication table", there are no obvious return variables for correctness checking, and providing an output formatting stub defeats the exercises' purposes. These strategies are "too invasive", as students are "told not only the existence of the [automatic] grader, but also the exact information to pass to it, and (in some instances) the exact point in their programs at which the call must be made" (Jackson, 1991). These strategies may have side-effects that disclose unintended hints to the expected form of solution, confining the way that students work on the problem (such as pre-defining the exact internal data structure) and limiting students' creativity (Tang et al., 2006). Finally, the driver/stub strategies require time-consuming effort in developing extra code specifically for every exercise.

3.6. Unit Testing Framework

Recent advances in the software testing field and test-driven software development have resulted in the popularity of generic unit testing frameworks, most of which are descendants or variants of JUnit. In these frameworks, a “test case” is implemented as a *test method*, which invokes the program unit under test and verifies the correctness of its actual return value(s) via `assert` statements (such as `assertEquals` or `assertTrue`), typically by comparing with its expected return value(s) included as parameter(s) of the `assert` statements. Many new automated systems, including Web-CAT (Edwards, 2003), AutoGrader (Helmick, 2007) and Oto (Tremblay et al., 2008), have incorporated the concepts or operation of JUnit or similar frameworks.

Figure 5 shows a JUnit test method for checking the time taken for travel, produced as return values computed by a student program for Exercise 1. Here students are required to implement the class `TimeCalcImpl`. The test method `testTime` passes the “input” values to `tc`, an object of the class `TimeCalcImpl`, by invoking the method `tc.setDistanceSpeed`. The test method then checks, via `assertEquals` statements, the values of the computed hour(s), minute(s) and second(s) returned by the methods `tc.getHour`, `tc.getMinute` and `tc.getSecond`, respectively. Such an implementation essentially employs the driver strategy. Thus, using a generic unit testing framework reduces, but not eliminates, the custom code written by the instructor.

```
import org.junit.*;
import static org.junit.Assert.*;
public class TimeCalcTest{
    @Test
    public void testTime(){
        TimeCalc tc = new TimeCalcImpl();
        tc.setDistanceSpeed(125, 90);    // set distance and average speed
        assertEquals(1, tc.getHour());   // check the return value of hour
        assertEquals(23, tc.getMinute()); // check the return value of minute
        assertEquals(20, tc.getSecond()); // check the return value of second
    }
    ...
}
```

Figure 5. A JUnit test method for testing a program solution for Exercise 1.

3.7. Use of Regular Expression

Prescribing highly precise specifications or checking return values actually try to avoid, rather than solve, the problem of output variation (Saikkonen et al., 2001). Drivers/stubs and unit testing frameworks test individual subprograms, not complete programs that students in introductory programming classes normally write. In fact, as long as some variations in outputs are allowed, the need for a non-trivial output comparison method remains. For example, Exercise 1 has not specified the exact format when the time taken is not an integer. Instructors often accept the number of seconds whether rounded to (1) an integer, (2) one or two decimal place(s), or (3) three significant figures. (An air-tight specification would include an undue amount of details to specify the exact format for the above and many other cases.) To allow for this flexibility, the code for handling these variations may be embedded in the driver, stub or `assert` statements, all of which need to be custom written by the instructor for every exercise. As such, a lightweight approach is desirable to simplify such tasks. One way is to use regular expressions. Figure 6 shows some regular expressions designed to match the program outputs for the sample input in Exercise 1. Note that the outputs of both **Prog-A** and **Prog-B** will then be judged to be correct accordingly, but the misspelt word in **Prog-C** is still considered unacceptable.

```
[Dd]istance\s*[:=]\s*125\s*(km|kilometer|kilometers|kilometre|kilometres)
[Aa]verage\s*[Ss]peed\s*[:=]\s*90\s*(km|kilometer|kilometers|kilometre|kilometres)/(h|hr|hour)
[Tt]ime\s*taken\s*[:=]\s*1\s*(h|hr|hrs|hour|hours)\s*
23\s*(m|min|mins|minute|minutes)\s*20(.\|.0|.00)\s*(s|sec|secs|second|seconds)
```

Figure 6. Regular expressions for determining the correctness of outputs in Exercise 1.

Regular expressions can be used as part of the output comparison module in an automated system, such as the “oracle” program in Ceilidh (Zin & Foxley, 1996), or with a unit testing framework, as in HoGG (Morris, 2003). Although regular expressions are simpler than program code, they can become quite clumsy as more variations are allowed. Moreover, writing regular expressions for every test case is still a tedious task.

3.8. Use of Parser Tools

Jackson (1991) proposes the use of the lightweight parser tools, *lex* and *yacc*, in Unix systems. Using *lex*, a file (lex script) is prepared containing the definition of lexical items in the actual output string. This file is automatically transformed by *lex* into a program that returns the tokens extracted from the actual output. Meanwhile, a file (*yacc* script) is also prepared that defines the actions to be taken when a token (returned by the *lex* produced program) is encountered. Together, a few lines of *lex* and *yacc* scripts are often adequate to specify and generate a pattern recognizer of much greater flexibility than with regular expressions alone. For details, an extended example can be found in (Jackson, 1991). However, even though writing these scripts may be relatively easy for people who are proficient in *lex* and *yacc*, others may not wish to learn these tools solely for the purpose of specifying output variations for automated program assessment systems. Thus, though powerful, this approach has apparently not been widely adopted.

3.9. The Token Pattern Approach

Recently, Tang et al. (2009a) have developed a token pattern approach. They propose to decompose the output string into groups of successive characters, called *tokens*, that represent meaningful pieces of information. A *token pattern* refers to a string of tokens automatically extracted from the expected output, each having a *type*, *value* and associated (default) *matching rule(s)*. Matching rules are the criteria for determining correctness when the token is compared with the actual output. If the instructor is not satisfied with the default rules associated with some of the tokens, he/she may fine-tune them via a GUI. As an illustration, Figure 7 depicts part of a token pattern that corresponds to the last line of the sample output in Exercise 1.

In Figure 7, each token is shown as a rounded rectangular box. The first token has type “character” and value “time”. The rule “Ignore case” means that matching is case-insensitive, and “Correction” means that small deviations (such as minor spelling errors) are tolerated subject to automatic correction based on a built-in dictionary. The second token has type “whitespace” and is “Ignored” during matching as long as at least 1 whitespace character exists in the actual output token. The second token in the second line has type “character”, and its value can be any element of an external list named **H_LIST**, which contains all acceptable labels for the unit **hour** (such as **h**, **hr** or **hours**). Thus, matching succeeds when the actual output has any such acceptable label as value, after ignoring case sensitivity and applying dictionary-based correction. The first token in the third line has type “double” (double precision) and a value of 20.0. Matching succeeds with any number equal to 20.0, up to a “Precision [of] 2 d.p. (decimal places)”. Other tokens may be similarly interpreted.

Type	character	whitespace	character	whitespace	punctuation	whitespace	Integer
Value	time	Space qty: 1	taken	Space qty: 1	=	Space qty: 1	1
Rule(s)	Ignore case Correction	Ignore	Ignore case Correction	Ignore	Ignore	Ignore	
	whitespace	character	whitespace	Integer	whitespace	character	whitespace
	Space qty: 1	H_LIST*	Space qty: 1	23	Space qty: 1	M_LIST*	Space qty: 1
	Ignore	Ignore case Correction	Ignore		Ignore	Ignore case Correction	Ignore
	double	whitespace	character	whitespace	whitespace	*H_LIST, M_LIST, S_LIST are lists of acceptable labels for the units hour , minute and second , respectively.	
	20.0	Space qty: 1	S_LIST*	LF	CR		
	Precision 2 d.p.	Ignore	Ignore case Correction	Ignore	Ignore		

Figure 7. A partial token pattern derived from the last line of the sample output in Exercise 1.

The token pattern approach is flexible as it allows fine-grained matching rules for each token by the instructor using a GUI, without writing any programs or scripts. Currently, a prototype is under development, and preparation of the evaluation of its practical effectiveness is underway (Tang et al., 2010).

4. Summary and Conclusion

We have described the output correctness determination problem in automated program assessment systems. Using Exercise 1 in Figure 1 as a running example, we have systematically reviewed the various strategies used in practice to address the problem. In summary, the use of basic character matching and ad hoc character filtering or conversion strategies alone are unsatisfactory, and have been shown to bring about undesirable pedagogical issues such as student frustration and confusion (Jackson, 1991; Tang et al., 2009b). An adequate amount of precision and details in program specifications are certainly necessary, but prescribing excessive precision and details in an attempt to produce a unique correct output for each input is shown to be problematic.

Many instructors avoid the output variation problem by designing custom drivers/stubs, or by using a unit testing framework (such as JUnit). These strategies work fine for some types of programming exercises, but in other situations, they are considered too invasive, as explained in Section 3.5. More importantly, as long as non-trivial output variations are tolerated, the correctness determination problem remains. Some systems make use of regular expressions or parser tools, but again they require the instructor to write a script for every test case, which remains a tedious task even though writing these scripts may be easier than coding programs, drivers or stubs. Recently, a token pattern approach has been newly proposed, which promises to allow instructors to select, via a GUI, fine-grained comparison criteria for each output token without writing code or scripts. The approach is still under development, and further work is necessary to evaluate its potentials.

In practice, instructors usually use not just one strategy, but a combination of strategies. For example, some systems adopt unit testing frameworks and at the same time allow the use of regular expressions, and the instructor may also specify adequate details of program specification to avoid misunderstanding. Currently, instructors still spend great effort in dealing with the output correctness determination problem. Hopefully, the undesirable pedagogical consequences due to the problem, as well as the effort spent by instructors, can be significantly reduced with the adoption of the new token pattern approach.

References

- Ala-Mutka, K. (2005). A survey of automated assessment approaches for programming assignments. *Computer Science Education*, 15(2), pp. 83–102.
- Curator (2009). Curator: An electronic submission management environment. Retrieved December 12, 2009, from <http://courses.cs.vt.edu/curator/>
- Edwards, S. H. (2003). Improving student performance by evaluating how well students test their own programs. *ACM Journal of Educational Resources in Computing*, 3(3), Article 1.
- English, J. (2004). Automatic assessment of GUI programs using JEWL. In *Proceedings of the 9th Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE '2004)*, pp. 137–141.
- Harris, J. A., Adams E. S., & Harris, N. L. (2004). Making program grading easier: but not totally automatic. *Journal of Computing Sciences in Colleges*, 20(1) 248–261.
- Helmick, M. T. (2007). Interface-based programming assignments and automated grading of Java programs. In *Proceedings of the 12th Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE'2007)*, pp. 63–67.
- Higgins, C., Symeonidis, P., & Tsintsifas, A. (2002). The marking system for CourseMaster. In *Proceedings of the 7th Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE '2002)*, pp. 46–50.
- Jackson, D. (1991). Using software tools to automate the assessment of student programs. *Computers and Education*, 17(2), 133–143.
- Joy, M., Griffiths, N., & Royatt, R. (2005). The BOSS online submission and assessment system. *ACM Journal on Educational Resources in Computing*, 5(3), Article 2.
- Luck, M., & Joy, M. (1999). A secure on-line submission system. *Software — Practice and Experience*, 29(8), 721–740.
- Morris, D. S. (2003). Automatic grading of student's programming assignments: An interactive process and suite of programs. In *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference (FIE2003)*, pp. S3F-1–6.
- Nazir, U., Poon, C.K., Yu, Y.T., & Choy, M. (2005). Automated assessment for improving the learning of computer programming: Potentials and challenges. In *Proceedings of the 9th Global Chinese Conference on Computers in Education (GCCCE 2005)*, pp. 634–639.
- Reek, K. A. (1989). The TRY system -or- how to avoid testing student programs. In *Proceedings of the 20th SIGCSE Technical Symposium on Computer Science Education (SIGCSE 1989)*, pp. 112–116.
- Saikkonen, R., Malmi, L., & Korhonen, A. (2001). Fully automatic assessment of programming exercises. In *Proceedings of the 6th Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE 2001)*, pp. 133–136.
- Suleman, H. (2008). Automatic marking with Sakai. In *Proceedings of Annual Conference of the South African Institute of Computer Scientists and Information Technologists 2008 (SAICSIT 2008)*, pp. 229–236.
- Shukla, R., Carrington, D., & Strooper, P. (2005). A passive test oracle using a component's API. In *Proceedings of the 12th Asia-Pacific Software Engineering Conference (APSEC 2005)*, pp. 561–567.
- Tang, C. M., Yu, Y. T., & Poon, C. K. (2009a). An approach towards automatic testing of student programs using token patterns. In *Proceedings of the 17th International Conference on Computers in Education (ICCE 2009)*, pp. 188–190.
- Tang, C. M., Yu, Y. T., & Poon, C. K. (2009b). Automated systems for testing student programs: Practical issues and requirements. In *Proceedings of the International Workshop on Strategies for Practical Integration of Emerging and Contemporary Technologies in Assessment and Learning (SPECIAL 2009)*, pp. 132–136.
- Tang, C. M., Yu, Y. T., & Poon, C. K. (2010). An experimental prototype for automatically testing student programs using token patterns. In *Proceedings of 2nd International Conference on Computer Supported Education (CSEDU2010)*.
- Tremblay, G., Guérin, F., Pons, A., & Salah, A. (2008). Oto, a generic and extensible tool for marking programming assignments. *Software — Practice & Experience*, 38(3), 307–333.
- Yu, Y. T., Choy, M. Y., & Poon, C. K. (2006). Experiences with PASS: Developing and using a Programming Assignment Assessment System. In *Proceedings of the 6th International Conference on Quality Software (QSIC 2006)*, pp. 360–365.
- Zin, A. M., & Foxley, E. (1996). The “oracle” program. Retrieved April 1, 2009, from <http://www.cs.nott.ac.uk/~ceilidh/papers/Oracle.html> .

以 QTI 為基礎之線上動態評量管理系統發展及實驗

The development of QTI-based dynamic assessment management system and its science experiment

賴阿福

臺北市立教育大學資訊科學系

電郵：lai@go.tmue.edu.tw

吳明行、陳志鴻

臺北市立教育大學資訊科學系

電郵：{wums, fehoun} @tp.edu.tw

【摘要】為協助教師瞭解學生發展中的潛能，提升其學習之成效，並達成各評量系統間試題交流與分享，本研究設計一套以 QTI 為基礎之線上動態評量系統。本系統可提供多媒體中介教材作為學生學習之鷹架，以期減少教師實施動態評量之時間及人力的負擔。本研究以臺北市某國小六年級四個班 121 名學生為對象，採用索羅門四組設計(Solomon four group design)來探討本系統對其學習國小自然與生活科技領域中「簡單機械概念」之效果。研究結果顯示應用本系統之實驗組學生可提升其學習成效，達到學習遷移及保留的效果，尤其是前測低成就的學生之學習成效尤為顯著。

【關鍵詞】QTI、動態評量、簡單機械、多媒體中介教材。

***Abstract:** In order to facilitate the teachers to understand the learner's developmental potential and promote the student's learning effect, this study employed the web technology to develop a QTI-based dynamic assessment management system (DAMS for short). To investigate the learning effect of dynamic assessment under DAMS, this study adopted the Solomon-four-group design, employed gradual prompting strategy, and conducted an experiment for learning simple mechanism concept. The subjects were 121 sixth graders. The results indicates that (1) the scores of simple mechanism concept of treatment group were better than that of control group significantly, and (2) the gain scores of low achievers in the treatment group were better than that of control group more significantly.*

Keywords: QTI, dynamic assessment, simple mechanism concept, multimedia intervention materials.

1. 緒論

1.1. 研究背景

本研究開發符合 QTI(Question and Test Interoperability Specification)標準的測驗系統，以期縮短教師教材開發之時程、減少開發成本並促進教材在各學習平台間流通分享。本系統導入漸進提示動態評量(graduated prompting assessment)方式來測得學生的發展中潛能，並以數位化多媒體中介教材提供其學習之鷹架，期能提升學生的學習成效及遷移效果，並能提供教師有關於學生概念學習之歷程。

1.2. 研究目的

本研究延續劉智維(2006)及黃聰欽(2007)發展之系統，旨在發展一套以 QTI 為基礎之線上動態評量管理系統，使得教師完成的試題能在不同的測驗平台間交換，以達到試題的可重用性。本系統並導入動態評量之功能，實際應用於國小自然與生活科技領域中「簡單機械」概念之學習，期能藉此了解學生的學習潛能和促進其學習遷移及保留成效。

2. 文獻探討

2.1. QTI

全球線上學習國際聯合機構(IMS Global learning consortium, IMS) 以 XML-based 作為資料交換的規格制定了 QTI 標準，目的是為了在不同的平台之間可以互相使用教材、追蹤學習者的進度、回報使用者的成效及交換學生的紀錄而發展出來的一種開放性規格(IMS Global Learning Consortium, Inc, 2006)。本系統所採用的試題格式遵循 QTI 2.1 草案的規範，使評量的試題能夠具有可交換性及再使用性，以便達到不同平台的題庫資源共享。

2.2. 動態評量

傳統的評量方式是以結果為導向，僅能測得學生目前的學習成就，無法反應其真實的智力。因此，在 Vygotsky 提出社會認知發展論，強調社會互動、「潛在發展水準」及「近側發展區(zone of proximal development)」等觀點後，各學者便依此觀點發展出動態評量的理論。在各式動態評量模式中，Campione 和 Brown(1987)提出漸近提示評量，其以「前測—學習—遷移—後測」四個程序進行。前測目的在於了解學習者目前的學習表現；透過標準化的中介提示系統可分別偵測學習量數及遷移數；再測驗目的即在於了解學習的增進狀況；中介教學提示的順序乃由抽象到具體，評量時學習者若答錯時，施測者按照提示順序給予學習者提示，最後計分就依提示量多少來給分，提示愈多表示能力越差，分數也就愈低。漸近提示評量之提示系統經過事先分析排列、施測簡便，所以普遍被採用於學科上(賴阿福，2005)。

動態評量實際應用在學科上有許多的優勢，可有效區辨不同程度的受試者在「學習能力」以及「遷移能力」上的差異，利用多媒體動態評量模式則更能增強其學習情意(莊麗娟等，2001)。本系統結合動態評量功能，導入數位化多媒體中介教材，期能提升學生學習及保留的成效，教師並可以利用資料庫內所紀錄的學生答題歷程進行其學習概念之分析。

3. 研究方法

3.1. 研究對象與研究設計

本研究以臺北市某國小六年級四個班 121 位學生為研究對象，採用所羅門四組準實驗設計方式進行學生利用本系統學習的成效之探討。首先，針對實驗甲組及控制甲組學生，以學習成就測驗甲卷進行前測，之後實驗甲、乙組分別於電腦教室進行兩次線上動態評量，最後實驗組與控制組均以學習成就測驗乙卷進行後測。四週後實驗組及控制組學生均再以學習成就測驗乙卷進行延宕測，以了解學生在經過本系統實驗處理之後的學習保留成效。

3.2. 研究工具

本研究使用之工具包含 1.伺服器端電腦、系統開發軟體和用戶端電腦：開發系統之程式語言為 ASP.NET 2.0，資料庫為 Microsoft SQL Server 2005；2.成就測驗甲、乙卷：甲卷為前測測驗卷，試卷難度為 .58、鑑別度為 .65、庫李信度為 .91；乙卷為後測及延宕測測驗卷，試卷難度為 .57、鑑別度為 .65、庫李信度為 .89；3.多媒體數位化中介教材；4.系統評估問卷。

3.3. 系統架構及功能

本系統包含概念管理、題庫管理、測驗管理、診斷模組等模組。在動態評量功能上，本研究導入之動態評量採用漸進提示方式，其過程是由顯示題目開始，學生答對便進入下一題，答錯則顯示提示，直到提示終止時(如圖 1a 至圖 1d)。

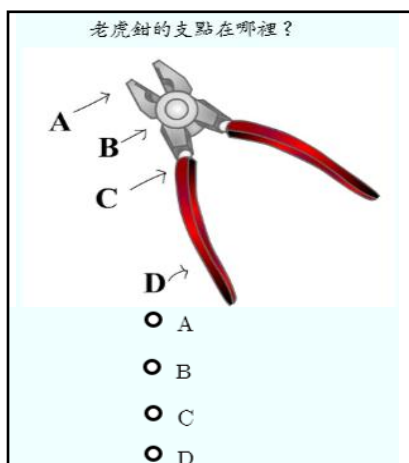


圖 1a 動態評量單選題測驗畫面



圖 1b 動態評量第一次提示

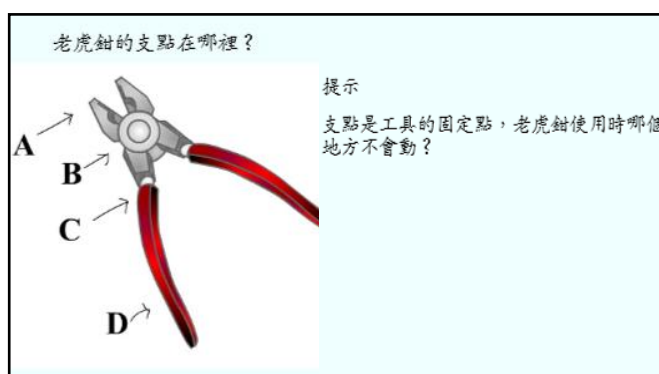


圖 1c 動態評量第二次提示



圖 1d 動態評量最後直接教學

本系統之動態評量提示方式分成下列四個階段進行：1.直接回答正確，不需提示；2.提示一：題意協助。提供題意理解的協助，讓學童可以回憶其先備的解題知識來幫助解題；3.提示二：關鍵提示。提供關鍵(字)的提示，來給予題意結構解析的支持；4.提示三：直接教學，運用 Flash 動畫式教材提供視覺化學習。

4.研究結果與討論

4.1.系統和數位化試題及中介教材評估

本研究在系統所提供之功能對於教師之操作及助益進行評估，分析結果顯示其平均得分為 4.59(滿分 5 分)；Cronbach $\alpha = .732$ 。評估者大多認為本系統容易編製試題、製作測驗、建立概念結構，以及藉由系統提供的資訊可了解學生的學習歷程，能促使學生概念的釐清與教師教學目標的改進，而依循 QTI 標準有助於未來試題的分享與簡化測驗的編製；在漸進提示動態評量數位化試題及中介教材設計方面，分析結果顯示數位化試題及中介教材向度的 Cronbach α 值為 .88；平均得分為 4.45 分。其中，值得注意的是在教材觀念表達方面獲得 4.52 分，顯示評估者認為本系統提供之動畫能將教材觀念表達清楚；在中介提示之教材設計方面獲得 4.41 分，顯示評估者認為此透過不同中介階層的動態圖像的優點能提供學生學習時所需之鷹架。

4.2.學生學習及保留成效

本研究以線上動態評量實驗處理（有無） \times 前測（有無）進行雙因子變異數分析，以考驗整個實驗組（甲、乙）和控制組（甲、乙）後測成績間是否有顯著差異存在。經資料統計分析後，得到實驗處理和前測之間交互作用 P 值為 $.059 > .05$ 並未達顯著性；而是否經過線上動態評量實驗處理之 P 值為 $.001 < .05$ 已達顯著水準，由此可知經過線上動態評量實驗處理後的學生在後測成績上明顯高於控制組學生，而是否接受前測對於後測成績則無顯著性影響，前測與實驗處理間的交互作用亦未顯著。

在延宕測驗結果分析上，控制組平均得分為 68.95 分，實驗組平均得分為 82.20 分， P 值為 $.008 < .05$ 已達顯著差異。由此得知，實驗組學生在經過線上動態評量學習後，在延宕測驗分數明顯高於控制組學生之分數，即實驗組學生在學習保留成效上比控制組學生較為顯著。

本研究進一步將低學習成就組學生以實驗處理為自變數，前後測分數差異為依變數，進行單因子變異數分析。經資料統計結果，低學習成就控制組學生之後測成績較前測成績退步 0.56 分；實驗組學生則進步 24.09 分， P 值為 $.016 < .05$ 已達顯著差異，亦即低學習成就實驗組學生在經過線上動態評量學習後，後測成績進步量高於控制組學生之後測成績進步量。莊麗娟（2001）提出低獲益者有三方向的認知缺陷：1. 缺乏結構性思考；2. 工作記憶不足；3. 遷移力缺陷，提供結構運算是強化思考方向，避免思考散亂並減輕工作記憶負荷的協助策略，本研究以多媒體方式提供各層次之中介提示，所得結果與其一致。

5. 結論與建議

本研究藉由 QTI 2.1 標準將試題資源標準化及概念結構化，便於教師在不同平台上建立及管理以概念為基礎的線上題庫，達到測驗資源共享之目的。相較於先前國內 QTI 系統的研究，本系統著重於回饋型試題的新增，並導入漸進提示動態評量之功能。教學實驗研究結果顯示，經由本系統之線上動態評量可開發學生之學習潛能，提升學習成效及促進學習遷移並具保留成效，其中低學習成就學生尤為顯著。如何提升低學習成就學生之學習成效為教育者所關注之課題，未來可針對多媒體中介教材之線上動態評量方式加以進一步研究。

參考文獻

- 莊麗娟（2001）。「多媒體動態評量」低獲益受試者之認知缺陷與協助策略分析。《特殊教育研究學刊》，第 21 期，109-133 頁。
- 莊麗娟、邱上真、江新合、謝季宏、羅寶田（2001）。多媒體動態評量模式之效益分析—以自然科學「浮力概念」為例。《中國測驗學會測驗年刊》，第 48 輯第 1 期，43-70 頁。
- 賴阿福（2005）。電腦化動態評量系統與多媒體中介教材設計研究。2005 年臺北市資訊教育國際研討會論文集(2005 International Conference on ICT in Education)，121-134。
- Campione, J. C., & Brown, A. L. (1987). Linking dynamic assessment with school achievement. In C. S. Lidz(Ed.), *Dynamic assessment: An interaction approach to evaluation learning potential* (pp. 82-115). New York: Guilford Press.
- IMS Global Learning Consortium, Inc. (2006). *IMS Question and Test Interoperability Overview Version 2.1 Public Draft (revision 2) Specification*.

電腦化動作技能測驗系統之發展與驗證

The Development and validation of the Computerized Motor Skill Testing System

蕭顯勝

臺灣師範大學 科技應用與人力資源發展學系

郵件信箱：hssiu@ntnu.edu.tw

宋曜廷

臺灣師範大學 教育心理與輔導學系

郵件信箱：sungtc@ntnu.edu.tw

林建佑

臺灣師範大學 科技應用與人力資源發展學系

郵件信箱：897710024@ntnu.edu.tw

邱敬尊

臺灣師範大學 科技應用與人力資源發展學系

郵件信箱：697710480@ntnu.edu.tw

【摘要】本研究發展電腦化動作技能測驗系統，以作為國中生升學參考工具。利用電腦化測驗之優勢使測驗資料更容易蒐集與分析。本研究參考現有動作技能測驗計分方式設計試題內容與測驗流程進行試題設計與參考，並針對所參考之測驗進行測驗等價分析。受試者測驗介面使用 Wii remote 手把偵測受試者動作以控制畫面物件。系統完成後進行信度與效度分析，建置測驗常模，以期能對國中生有所助益。

【關鍵字】 動作技能、電腦化測驗

***Abstract:** This study developed a computerized motor skill test system. After finish the system, the system will be tested the reliability and the validity, and will be compared with the motor skills tests of the GATB test. Finally will create a system norm. The motor skill testing system use the Wii remotes as testing interface. The Wii remotes are able to detect human movement, especially the hand movements, and send data immediately into computer. Thus, system could analyze subject's motor skills. The computerized motor skill test system will be used as a junior high school student's careers guidance tool. In the hope that can be helpful to junior high school students.*

Keywords: Motor skills, Computerized test

1.前言

目前台灣實施的學制為九年一貫國民教育，九年級學生在畢業後立即面臨高中職入學的教育分流的問題。在多元入學的環境下，學校提供各種職涯測驗與性向測驗等心理測驗幫助學生作為選擇未來目標的參考依據。然而這些測驗許多下列問題，如：題目內容不合時宜、性向測驗包含之能力向度不符需求、施測不易等缺點（何榮桂，2000）。而這些性向測驗中測量

動作技能之動作技能測驗並不多，不容易對應到現行職業類群，實際進行施測需要耗費許多人力與準備器材，因此學校甚少為學生準備動作技能測驗。本研究初步調查發現高職 15 學群中有 11 學群相當重視手部動作技能，且各學群所重視手部動作技能亦不盡相同，如：機械群較重視手指靈巧；海事群較重視手部靈巧；餐飲群與家政群同時重視手指與手部靈巧。因此發展動作技能測驗用於學生性向分析是有其必要性。

本研究以發展一套電腦化動作技能測驗系統，研究成果將做為九年級學生升高中職時職涯技能分析的參考工具。希望以電腦化測驗的長處，如施測容易、資料蒐集方便、降低人力、減少施測成本（林敏芳，2005）等，結合體感技術與電腦化評分機制進行系統開發，讓系統方便學校與學生使用。系統開發完成後將進行預試與專家分析作為系統修正參考依據，並分析系統與現有之實作技能測驗是否等效。未來將進行實際測試來蒐集資料，以分析信度與效度，並建立測試系統常模，以利後續推廣與研究。

2. 文獻探討

2.1 動作技能測驗

動作技能(Motor skills)是為了達到某個目標而能精確的使用身體的主要動作(Magill, 2003)。為測量受試者控制動作的能力，因此發展動作技能測驗。根據測驗目的可將動作技能測驗分為兩類，分別是受試者適性測驗及受試者動作障礙測驗。受試者適性測驗為測量受試者能力是否適合、有能力進行某種動作技能，如體育訓練甄選、員工篩選、學生性向測驗等。而受試者動作障礙測驗為測量受試者能力動作障礙以提供適當的輔導。

本研究所參考之通用性向測驗為職訓局所發展之可通用於各種職業的性向測驗組合。通用性向測驗可以用於人員甄選與職業諮詢，也常做為國內學生性向測驗時所使用之工具。本研究調查行政院勞工委員會職業訓練局網站發現通用性向測驗目前仍常作為人員甄選或就業安置工具，因此本研究選擇通用性向測驗做為發展依據。

2.2 電腦化動作技能測驗之現況

由於動作技能測驗需要使用特別的工具，目前標準化電腦化動作技能測驗並不多。普渡大學發展之普渡釘板測驗(Purdue pegboard)(Lafayette Instrument Company, 2002)，主要是評估手及手指的靈巧度。這個測驗要求受測者在 30 秒內分別用右手、左手和雙手把釘子插到孔中，計算平均完成數。其電腦軟體版本，動作技能測驗軟體可提供方便的測驗環境，及包含自動計分之功能，可作為大量施測之使用。受試者操作滑鼠將每個釘子放置在釘板中，測驗後系統即可提供受試者的測驗分數；此軟體之發展證明利用電腦作動作技能之測試是可行的。

2.3 動作擷取技術

本研究調查人體動作擷取技術將擷取技術分成兩大類，分別是非接觸類與接觸類。非接觸類是藉由機械讀取人體運動時的反射波，以非接觸的方式取得人體資料。接觸類為利用物理裝置以接觸的方式偵測人體動作。近來由於三軸加速度晶片、陀螺儀晶片等電子元件量產，物理感應裝置小型化且成本降低，使得應用物理感應人體動作的限制減少許多。

本研究擬採用之 Wii remote 手把同時使用紅外線定位與加速度晶片，互相彌補技術缺陷以取得較精確、複雜之人體動作。Wii remote 手把利用放置於螢幕側紅外線發射裝置取得手把相對於螢幕之空間座標；利用加速度晶片取得手把運動資料。並經由藍芽裝置將所取得資料傳輸至電腦，藉由分析數據即可判斷手腕動作，再利用手把上按鈕偵測手指，藉由這些動作偵測能力作為測驗系統偵測手部運動之方式。本研究利用 Wii remote 手把與第三方廠商所開發之手把控制 API，自行應用 C#設計測驗系統做為施測工具。

3. 電腦化動作技能測驗系統

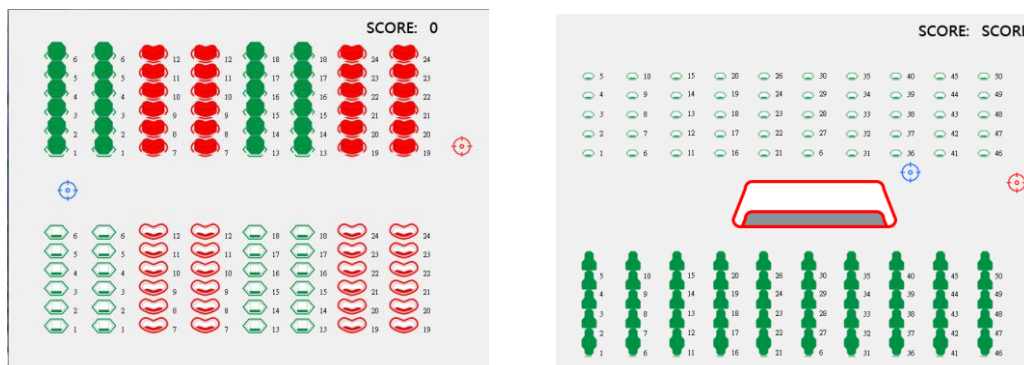
本研究參考通用性向測驗（中國測驗學會，1985）建置一套電腦化動作技能測驗系統，利用電腦化測驗之優勢，一方面可在最低成本下大量進行施測，另一方面則可利用電腦強大的運算能力，快速統計受測者的成績。系統完成後將進行電腦化動作技能測驗系統與傳統動作技能測驗之等價驗證。

3.1. 通用性向測驗動作技能分測驗

本研究之測驗系統試題設計主要參考通用性向測驗之分測驗 9 至 12 動作技能測驗之施測與計分方式。分別是拆開與組合，此測驗為測驗手指靈巧度，靈活運用手指將小物品拆開與組合之能力；移置與轉置，此測驗為手部靈巧度，靈活運用手腕、手肘與手將物體快速、精確的移動或轉動之能力。

3.2. 電腦化動作技能測驗

系統將使用 C# 語言進行設計之電腦化動作測驗系統，施測畫面如圖像 1，根據高職學群動作技能需求調查發展測驗試題內容，並參考通用性向測驗之分測驗 9 至 12 測驗流程與測驗時數進行設計。手指靈巧度測驗分為 2 個子測驗，分別是移置測驗與轉置測驗，計分方式與時間限制參考自通用性向測驗。本測驗分別將畫面所示之物品移置或轉置至指定位置，移置測驗雙手持 Wii remote 手把進行；轉置測驗使用慣用手持 Wii remote 手把進行。手部靈巧度測驗分為 2 個子測驗，分別是組合測驗與拆開測驗，計分方式與時間限制參考自通用性向測驗。本測驗分別將畫面所示之物品組合或拆開至指定位置，兩測驗皆雙手持 Wii remote 手把進行。



圖像 1 施測畫面

4. 系統驗證

本研究建置一套電腦化動作技能測驗系統，本研究之系統做為九年級學生升高中職時職涯技能分析中動作技能分析的參考工具。希望以電腦化測驗的長處，讓學校與學生更方便使用。系統完成之後進行下列三項工作，分別是系統信效度驗證、系統等價驗證與正式施測。各項工作細節描述如下：

4.1 系統信效度驗證

信效度驗證施測人數約國中九年級學生 500 人。測驗資料分析、修題，根據測驗資料進行統計分析，針對分析結果對試題進行修訂。系統驗證完成以及預試結果確認系統沒有使用問題之後進行實際施測，施測對象為國中九年級。實際施測所取得的數據進行信度與效度分析，做為判斷系統可信度與準確度之依據。效度資料將邀請 5 位動作技能專家與電腦化測驗專家，以分析系統測驗向度與測驗內容是否相符合之內容關聯效度。測驗信度根據測驗特性採重測

信度，分析重測信度時，測驗間隔時間越短則信度越高，但須注意受試者練習所造成之影響（郭玉生，2004）。

4.2 系統等價驗證

為證明電腦化動作技能測驗系統之有效性，將進行電腦化動作技能測驗系統與通用性向測驗動作技能測驗之等價驗證。等價驗證在測驗上的意義為表示傳統測驗與電腦化測驗兩種型式並存的可能性。等價驗證必須要證明經驗等價及統計等價兩項資料。另外，統計等價是指受試者在傳統測驗與電腦化測驗之兩個得分必須是一致的。本研究將以專家會議方式確認電腦化測驗之經驗等價；以預試方式蒐集受試者的傳統測驗與電腦化測驗兩種分數，作為系統統計等價驗證之用。

4.3 施測

等價驗證與信度效度驗證之後建立系統測試常模。系統測試常模用於分析受測者的能力位於常模參照的哪個位置，意即受試者能力與其他多數人相比大致如何。受試者得到自己的能力指標之後，可用於職涯決策與未來進學的參考資料。

5. 結論

九年級學生未來將面對選擇高中或高職的重要人生選擇，然而目前性向測驗大多只針對高中部分，對於高職部分所需大都沒有設計出適合的性向測驗作為學生選擇的參考依據。本研究以動作技能為著眼點，試發展一套作為九年級學生職涯選擇的決策工具，以期未來系統發展成功，能讓學生對於自己的能力與現實的職業做配合測試，進而做出最適合自己的決定。

誌謝

本研究承蒙行政院國家科學委員會專題研究計畫(計畫編號 98-2511-S-003-024-MY3, 97-2511-S-003-010-MY3, 98-2511-S-003-033-MY3, 98-2631-S-003-005-)補助經費，特此致謝。

參考文獻

- 中國測驗學會（1985）。《通用性向測驗指導手冊》。台北：行政院勞委會職業訓練局。
- 何榮桂（2000）。遠距測驗與評量。《2000 網路學習理論與實務研討會論文集》，新竹市。
- 林敏芳（2005）。線上評量應用於教學上的現狀與發展。《生活科技教育月刊》，38（1），74-85。
- 郭玉生（2004）。《教育測驗與評量》。台北：精華。
- Lafayette Instrument Company(2002). *PURDUE PEGBOARD TEST USER INSTRUCTIONS*. Lafayette: Lafayette Instrument Company.
- Magill, R. A. (2003). *Motor learning and control: Concepts and applications* (7th ed.). New York: McGraw-Hill.

學習歷程檔案評量研究之發展與趨勢分析

An Analysis and Trend of the Research of Learning Portfolio Assessment

劉力君、劉旨峰*

中央大學學習與教學研究所

pumpkinmay@gmail.com, totem@cc.ncu.edu.tw*

【摘要】本報告以 1995 年至 2008 年間所發表的 162 篇學習歷程檔案評量相關學位論文為主要樣本，目的在於瞭解台灣近年來在學習歷程檔案評量之研究方向，以歸結出現階段研究之成果。分析的編碼架構為以下幾個面向：「發表年份」、「紙本與數位化發表年份」、「研究者隸屬領域」、「研究對象」、透過編碼架構對於文獻資料進行分析。分析結果則提出在學習歷程檔案未來研究方向供研究者參考。

【關鍵字】學習歷程檔案、檔案評量、文獻分析

Abstract: This study surveyed 162 theses related to Learning Portfolio Assessment in electronic dissertations and theses system in Taiwan from 1995 to 2008 to identify research topics of this field and concluded the current status of this field. The coding schemes developed in this study include: publish year, field, topic, educational level, subject, Content analysis was applied in this study, and the coding structure was used to analyze the data. Finally, the discussion and trend was proposed to discuss the research process of Learning Portfolio Assessment in Taiwan

Keywords: Learning portfolios、Portfolios Assessment、Literature Analysis

1.前言

學習的過程是豐富且有機的，紙筆評量只能看到學生單向的能力展現，而近年來興起的學習歷程評量，則納入多樣的呈現媒介和學習者產出，替評量工具找到另一種可能性，隨著新科技發展帶入教育領域，紙本的學習歷程評量逐漸結合數位技術與管理平台，成為一個不可或缺的教學應用工具，國內從 1995 年起有研究者針對學習歷程評量主題撰寫論文，近年來也逐步累積出豐富的研究文獻，研究者對目前台灣地區碩博士學位論文在學習歷程評量這個主題下的研究主題與趨勢感到好奇，希望能透過分析論文的發表年份、隸屬學校、領域、研究對象、研究方法及研究主題等面向，了解論文整體面貌以達溫故知新之效，同時推測學習歷程評量未來的發展方向。

2.文獻探討

2.1 電子化學習歷程檔案（Electronic Portfolio）的興起

傳統學習歷程檔案的建立，仍有賴於人為的蒐集與建立以紙筆為主的學習過程資料，其長期大規模實行後所造成的資料儲存及搜尋、管理上的困難，這些以紙本或卷宗為基礎的檔案雖然適合應用在學習評量與教學評鑑上，但是當檔案數量越來越多，內容型態越來越多樣時，必然造成檔案收納、儲存及管理的問題，可能有毀損或遺失的風險，也缺乏即時更新管理的便利性，在查詢與交流上更受限於實體卷冊，難以將檔案加值作更有效的利用（Georgi & Crowe, 1998）。

為了解決舊有學習歷程檔案在資料儲存、管理、分享與格式上的諸多限制，將資訊科技

應用於學習歷程檔案，改用數位化方式儲存與發表，可使得學習歷程檔案具有體積小、不佔空間、輕便易攜帶及易保存等長處（岳修平、王郁青，2000），這就是近幾年受到越來越多研究者關注的電子化學習歷程檔案（Electronic Portfolio）。

3. 研究方法

3.1 研究樣本

學習歷程檔案評量屬於實作評量（performance Assessment）的一種，此評量的概念興起於 1960、1970 年代的美國測驗學界，到了 1980 年代末期正式且有系統地應用在教育領域上（Arter & Spandel, 1992），在美國已經有許多州或學區將學習歷程檔案作為變通評量（alternative assessment）的方式（Vavrus, 1990），在台灣最早與檔案評量有關的論文出現於 1995 年，承接國外發展多年的學理基礎，並與教育當局推動終身學習的教育理念相呼應，經界定後本研究的選擇時間樣本介於 1995 年至 2008 年，共 14 年間撰寫發表的學位論文作為分析的樣本來源，本研究主要以全國博碩士論文資訊網及 CETD 電子學位論文服務網作為樣本資料蒐集的檢索平台，。

3.2 研究流程

研究者首先針對主題進行關鍵字搜尋，所使用的關鍵字，中文部分包含卷宗評量、學習歷程評量、檔案評量，英文部分包含用關鍵字：Portfolio assessment、Learning portfolio、Learning profile、E-Portfolio，初步蒐集到的資料共有 178 筆，最後共有 162 篇論文資料作為本次研究的分析文本來源，其中博士論文 7 篇，碩士論文 155 篇，整理建檔後以內容分析編碼表進行歸類與統計，以呈現近 14 年來台灣地區碩博士生在學習歷程檔案評量研究的面貌，並進一步分析本研究領域的發展現況，並從資料佐證中推測未來研究趨勢。

4. 資料分析

4.1 學習歷程評量論文分析

4.1.1 研究發表年份

本研究所蒐集的範圍為 1995 至 2008 年，1995 年第一篇與此評量方式有關的學位論文題目為：「溶解刻板印象：兩性角色課程對國小學生性別刻板印象的影響」（劉淑雯，1995），研究者在課程實施後以統計及卷宗評量（Portfolio Assessment）方式了解學生在認知上的改變歷程，僅將其作為課程評量方法以評估教學成效並無深入探討學習歷程評量的重要性與發展。1995 年至 1999 年，五年間論文產出總計 10 篇，皆是碩士論文，應用的學科範圍遍佈程式設計、英文寫作、特殊教育、自然科學、師資培育、社會科教育，實驗對象以國小學生為主，代表研究者已注意到日益受重視的檔案歷程評量，嘗試與學科教學與評量結合進行初探性研究，其中有兩篇論文以國小教師和師資培育生為對象（詹寶菁，1997；羅美惠，1998），代表研究發展初期已有研究者關注到「教師專業成長」的主題，Sung 與 Chang（2004）亦指出近十年來，歷程檔案在師資培育機構和教師在職進修的環境中的應用受到許多研究者提倡，許多師資培育機構採用歷程檔案來作為提升教師專業發展的教學方法。許多研究者也發現透過教學卷宗的應用，可以促進教師在學科或教學知識、教學實務、學習過程與自我反省等方面的進步（Darling-Hammond & Snyder, 2000），因此未來可針對如何將學習歷程檔案用於師資培育與教師專業發展，擬定進一步的落實方針。

2000 年後相關研究論文發表數量逐年成長，2002 年至 2008 年每年皆有 16-23 篇的穩定發表量，代表有固定的研究社群持續關注學習歷程評量議題，並逐年注入不同的關注焦點（例

如與多元智能、建構式教學、幼兒教成長歷程、教師評鑑等議題結合)。

4.1.2 紙本學習歷程 檔案與數位化學習歷程檔案的發表篇數

Georgi 與 Crown (1998) 指出將紙本學習歷程檔案數位化後的優點有 1.產品資料的多樣化。2.方便的資訊儲存與管理。3.容易散佈與呈現資訊，且容易獲得回饋。由於應數位學習潮流及資訊技術的演進，如圖 1 統計圖表所示，自 2001 年起，有越來越多的研究者將學習歷程從傳統紙本卷宗檔案，改以數位化方式呈現運用 (5 篇)，到了 2003 年，與學習歷程檔案有關的論文運用數位 (9 篇) 和傳統紙本 (7 篇) 數量並進，且數位化檔案有持續升高超越紙本之趨勢。

4.1.3 研究者隸屬領域

從中可看出科學教育領域發表 32 篇論文明顯高於其他教育科目，推測其原因是導於科學教育牽涉到觀念的啟發、創造力及解決問題能力的培養、以及科技程序的教導，科學教育之教學評量自然不能只以單一評量方式達成(王應文,1996)，進一步細看科教領域的研究主題，無論是紙本或數位化學習歷程評量，都有研究者投入，教育層級以中小學為大宗，也應用於開設在大學的科技與社會學程 (STS)、高中數理資優班，除了進行教學成效評估、課程方案落實、將數位化學習歷程平台融入教學等全面性的研究方向，亦透過學習歷程分析開發出教學模組 (蔡曉信,2004)，讓更多實務教師可供運用，可見得若一領域研究者長期投入特定方法的理論建構與實證研究，將可開闢出特有的研究重心與優勢，建議其他教育學科的研究者也可循此模式找尋與學科精神扣合的評量方式加以深耕。

再者，數量次高的資訊工程與資訊管理領域各有 20 篇及 19 篇論文發表，研究方向以建置網路學習歷程檔案系統、資料探勘與管理工具研發為主。Wolf (1989) 認為建立電子化的作品集是一件非常困難及耗時間的事，也不是所有的學生都能夠輕易的撰寫網頁程式，所以勢必需要一個可以讓學生方便建立作品集的工具，因此資訊專業背景研究者的持續研發，是未來數位化學習歷程檔案推廣讓更多學生易於參與的必要推力。

此外，除了特定學科教育領域外，幼兒教育、技職教育、藝術教育與成人教育領域也有研究者運用學習歷程檔案進行學習評量以促進專業人員與學習者生涯發展，學習文化的塑造與發展，不但要推動終身學習，還必需要發展適宜的學習成就認證與評量方法，2006年起澳洲政府委託大學終身學習中心推動的Australian ePortfolio Project (AeP) 就利用數位化歷程檔案來支援終身學習的推廣、記錄與認證，見證學習者終身學習的歷程 (McAllister, Hallam & Harpur,2008)，雖然目前有關非正規教育領域的研究論文篇數較少，且為最近五年新發表，研究重心尚未聚焦，但由於社會變革所趨，已隱然成為未來可期的方向，因此如何用運用學習歷程檔案鼓勵與支持全民終身學習的推廣，是未來研究者可嘗試深入探究的途徑。

4.1.4 研究對象

在研究對象部分，有高達 80 篇以國中小學學生為研究對象，其次依序為沒有標明特定對象的 29 篇，研究的主題多偏向系統功能開發、大學學生 16 篇、高中職學生 15 篇、中小學老師 7 篇、幼稚園老師 5 篇。石維婷 (2005) 指出目前學習歷程檔目前在北美、歐洲已受到普遍支持，著重在應用、推廣階段，而台灣、中國目前則為實驗階段中，實驗對象多集中在中、小學。亦有研究者指出當新的研究議題剛出現的時候，研究的取向會針對具有普遍特徵的對象 (例如中小學生) 進行初探研究，隨著研究的增加以及對該議題的了解，會轉換至具有特殊需求的研究對象上。

在沒有標明特定對象的 29 篇中多屬於系統平台開發與功能設計之主題，研究歷程大多止於系統實作，並無卻無進一步安排問卷與實驗設計和使用者的測試，確認在特定情境運作時可

能發生的非預期問題，在進一步對照研究者領域後發現多屬資訊系所研究者，他們具有支援教學的技術能力，但和教學現場並無聯繫管道，為了避免過度重視系統功能忽略使用者需求的缺憾存在，建議未來可搭建聯繫管道讓教育現場的老師能與開發系統的研究者協同研究，讓進行數位化學習歷程評量時仰賴的平台和工具能更貼近實務操作之用。

再者，以老師作為研究對象的論文篇數不多（中小學老師7篇、幼稚園老師5篇、大學老師1篇），Freeman（1998）曾以檔案做為教師專業成長工具進行個案研究，發現檔案可以促進教師自我反省、自我評鑑，並提升專業成長的動機，建議研究者可針對教師對歷程檔案的認知、接受度進行調查研究，在對學生進行學習歷程評量時同時納入老師的教學歷程。此外，近幾年由於大學法第二十一條規定大學必須定期進行教師評鑑，各大學紛紛訂定教師績效評量辦法，而教學歷程檔案是一有效的評鑑資料參考，建議未來可針對各層級的教師教學歷程檔案作進行意向調查、需求評估、與系統建置研究。

5. 結論與未來建議

根據本文分析發現，在研究主題方面，未來可針對如何將學習歷程檔案用於師資培育與教師專業發展，擬定進一步的落實方針。在研究對象方面建議未來可針對各層級的教師教學歷程檔案作進行意向調查、需求評估、與系統建置研究。在研究者背景方面，若有更多普通大學及技職院校研究者從多元背景注入活水進行更廣泛的討論與對話，將有助於理論與實踐之更新。在研究方法方面，建議可針對歷程評量採用多樣研究方法，將有助於增進理解與發現潛在問題。

此外，由於新科技與觀念的演進，可結合具備 web2.0 概念的平台，結合出新式學習歷程建置模式，並建立學生社會網絡社群（social network）促進學習歷程檔案的交流，加上學生面臨職場就業的競爭，可進一步把歷程檔案從課堂評量，轉變為輔助學生職涯發展、生涯規劃的評估資料，及專業能力的佐證。

謝誌

感謝國科會科教處對本研究的贊助，計畫編號為 NSC 97-2511-S-008-003-MY3

參考文獻

- 王應文（1996）。科學素養的評鑑。中學工藝教育月刊，29(6)。
- 岳修平，王郁青（2000）。電子化學習歷程檔案實施之態度研究。教育心理學報，31卷，2期
頁65-84。
- 卓宜青、劉旨峰、袁賢銘、林珊如（2000）。網路化學習歷程系統之初探性研究。新加坡大學
全球華人教育資訊科技學術研討會（GCCCE 2000）
- 張美玉（2000）。歷程檔案評量的理念與實施。科學教育月刊，231期。
- Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and Teacher Education*, 16, 523-545.
- Freeman, J. J.(1998). The teaching portfolio as a vehicle for professional growth (portfolio).PHD. Unpublished doctoral dissertation. University of New Hampshire
- Georgi, D. & Crowe, J. (Winter 1998). Digital Portfolios: A Confluence of Portfolio Assessment and Technology. *Teacher Education Quarterly*, 73-84.
- McAllister, L., Hallam, G., & Harper, W. (2008). The ePortfolio as a tool for lifelong learning: Contextualising Australian practice, *International Lifelong Learning Conference* (pp. 246-252). Yeppoon, Queensland.

線上閱讀動機量表編製

Developing a Motivation for Online Reading Scale (MORS)

賴怡君、張瑜芳、劉旨峰

國立中央大學學習與教學研究所

copyxee@gmail.com, yakinsky@gmail.com, totem@cc.ncu.edu.tw

【摘要】 本研究主要目的在編製一份具有信度與效度的線上閱讀動機量表 (Motivation for Online Reading Scale, MORS)，此問卷的編制基於閱讀動機問卷以及線上閱讀特性。本研究發展之線上閱讀動機量表共有 14 個向度，分別為社群交流、回饋、即時便利性、表現、推薦、滿足幻想、閱讀效能、順從、閱讀享受、逃避閱讀、認可、挑戰、反思以及好奇。

【關鍵字】 閱讀動機；線上閱讀動機；閱讀動機量表；線上閱讀動機量表

***Abstract:** The purpose of the research is to develop a motivation for online reading scale with validity and reliability. Based on the motivation for reading questionnaire (MRQ); and the characteristics of online reading, a motivation for online reading scale developed by this study consisted of 14 dimensions: community interaction, feedback, convenience, performance, recommendation, fantasy, reading efficacy, compliance, enjoyment, avoidance, recognition, challenge, reflection, and curiosity.*

Keywords: reading motivation, motivation for online reading, motivation for reading questionnaire, motivation for online reading scale

1.前言

在現今的時代，知識的獲得日益重要，快速且有效率獲得知識的方式就是閱讀，透過閱讀人們可以學習他人花費心力而獲得的體悟(洪蘭和曾志朗，2001)。此外由傳統的書籍、報章雜誌，到網路上的各式網站、電子報、電子書、部落格、線上百科都成為人們獲取知識的重要管道。研究指出這些透過網路蒐集資訊的方式，尤其在青少年族群中受到歡迎，青少年上網的主要目的也與線上閱讀息息相關，譬如資訊搜尋與休閒娛樂(陳俞霖，2003)。網路延伸的相關工具增加了閱讀資料快速取得、更新與搜尋的機會，例如 Wiki、blog、線上書櫃等等，青少年可以透過線上閱讀獲取新資訊進而與他人分享、討論，甚至自行創作於部落格，透過部落格的連結與他人建立關係(賴怡君、張瑜芳、陳峰毅和劉旨峰，2009)，這些都是線上閱讀的優點(陳俞霖，2003)。然而過去對於閱讀的研究，多以傳統閱讀為主，包括報紙、雜誌與書籍等等，加上線上與紙本閱讀有一定的重疊性，因此本研究將以傳統閱讀動機文獻為基礎，再加上線上閱讀特性的文獻，擬編制線上閱讀動機量表。

2.文獻探討

2.1.動機與閱讀動機

閱讀是學習和吸收知識不可或缺的媒介(張春興，1988)，學習各領域的知識必須藉由閱讀來達成，惟有具備閱讀動機及實際從事閱讀的行為，學生才能持續不斷主動學習其他新知識。Vygotsky (1978)認為閱讀不僅是被動的接受，更是主動的認知理解與意義，或與他人磋商互動的歷程，此種深度閱讀有利於知識的轉化與應用。動機是引起個體活動，維持已引起的活動，並促使該活動朝向某一目標進行的內在歷程(張春興，1988)，而「閱讀動機」則是指引起個

體的閱讀活動，維持已引起的閱讀活動，並促使該閱讀活動朝向某一目標進行的內在歷程。Wigfield, Guthrie, and McGough (1996)認為閱讀動機包括效能、挑戰、好奇、投入、重要性、認可、成績、競爭、社交、順從、逃避等成分，線上閱讀動機若加上網路特性，是否仍能獲得同樣的向度，是本研究探討的重點之一。

2.2. 閱讀動機量表

Wigfield 等人(Baker & Brown, 1984; Wigfield & Guthrie,1997)，根據閱讀動機相關理論、教室觀察資料、實徵資料及訪談，發展一套閱讀動機量表(The Motivation for Reading Questionnaire, MRQ)，將動機內涵分為三個面向11個要素，包括能力及效能信念、內在與外在動機，主觀價值，成就目標、以及社會因素。說明如下：

- **能力及效能信念：**指個體瞭解自己能否成功閱讀的能力及效能信念，成份包括：效能、挑戰以及避免閱讀工作。
- **成就價值及目標：**指閱讀者的個人目標，包含成就目標、工作價值與內、外在動機。建構在此向度的閱讀動機成份包括：好奇、投入、重要性、認可、成績、競爭。
- **社會因素：**指閱讀時與同儕、朋友或家人分享，或透過追求閱讀意義而能成為某社群一員的過程，成份包括：閱讀的社會理由、順從性閱讀、逃避。

宋曜廷、劉佩雲和簡馨瑩(2003)修訂Wigfield與Guthrie (1995)所建構的閱讀動機量表，以台北縣市國小五、六年級與國中一年級835名學生為研究對象，結果顯示此份中文閱讀動機在三面向及十一個分量表皆和Wigfield等人(1995)閱讀動機理念的多元構念一致。黃馨儀(2002)編製的國小學童閱讀動機量表亦是依據Wigfield等人(1995)所設計的閱讀動機量表，將十一個閱讀動機向度分為內在動機與外在動機兩個層面。閱讀效能、好奇、投入、挑戰、重要性、與逃避閱讀工作歸類為內在閱讀動機；為認可而讀、為成績而讀、為競爭而讀、為社交理由而讀、以及因順從而讀歸類為外在閱讀動機。

2.3. 閱讀與線上閱讀

蔡慧美(2005)指出網路閱讀是一種以數位或電子形式呈現內容的閱讀方式，可分為廣義與狹義，前者為舉凡閱讀數位或電子形式的內容皆屬之，譬如閱讀電子書、電子報、電子雜誌或期刊、討論群組、瀏覽網頁、E-mail、多媒體影音，或使用網路書店、BBS 等閱讀；後者則專指經由網路及個人電腦閱讀電子期刊/雜誌、電子報、網路小說或文學、電子書(不含單機版電子書)，以及從紙本數位化的閱讀素材之閱讀。本研究之線上閱讀採廣義定義，因過去研究主要關注的焦點在傳統閱讀動機的部分，對於線上閱讀動機的部分仍屬於起步的階段。Wigfield 等人 (1995)閱讀動機量表向度的制定是作者根據課堂觀察、訪談以及動機理論、實徵研究而得，研究對象為國小學童，因此有一些向度並不適合網路情境，因為網路是一個具有主動性的平台，上網閱讀可能是基於自身為獲取資訊、為休閒理由而閱讀，並非被動的閱讀模式。因此本研究參考 Schutte 與 Malouff (2007)的研究，將為成績而讀這個向度修改成「表現」，除了可以包含為成績而讀之外，還可包括其他不同的表現情境，因此或許比為成績而讀較為適切。另也加上一些閱讀的向度如反思、推薦與滿足幻想，再加上網路特性：回饋、分享、即時便利性、吸引力、社群等等。隨著網路科技的普及以及網路使用者的增加，本研究希望以閱讀動機問卷出發，編制一份線上閱讀問卷，希望未來能透過此一問卷了解網路使用者線上閱讀動機與線上閱讀行為之間的關係。

3. 研究方法

3.1. 研究對象

本研究以立意取樣挑選桃園縣私立完全中學，包含國中部與高中部。分別由高中、國中部各選一班做施測，共 78 人，其中有效樣本為 72 人，高中 35 人；國中 37 人。

3.2. 線上閱讀動機量表編製流程

本研究之線上閱讀動機量表，參考 Guthrie, McGough, & Wigfield (1994) 閱讀動機量表的架構，該量表將閱讀動機分為 11 個向度，分別為閱讀效能、閱讀挑戰、閱讀好奇、閱讀享受、閱讀重要性、順從、閱讀認可、為社交理由而讀、閱讀競爭、逃避閱讀，並將為成績而讀修改成表現。由於該量表的施測是應用於教室情境，本研究另加入一些也屬於閱讀的向度，如推薦、反思、滿足幻想(蔡慧美，2005)，和網路向度，如即時便利性、回饋、吸引力、社群、分享(江靜之，2003)，再參考同樣以 MRQ 為基礎的研究(Schutte & Malouff, 2007; Watkins & Coffey, 2004; 李素足，1998; 宋曜廷、劉佩雲和簡馨瑩，2003)擬定題目，再經由閱讀與網路科技專家檢閱修正，確立量表的專家效度，初步編制的問卷共有 19 個向度，共 123 題。在計分方面，採用 Likert 6 點量表填答形式並以紙筆填答。

4. 資料分析

4.1. 問卷刪題流程

本研究的刪題與分析的程序依序為描述統計量與試讀、信度分析 I、項目分析以及因素分析，最後是信度分析 II 等等，共計刪掉 53 題，剩下 70 題。

4.2. 因素分析

本研究採主成份法 (Principal Component) 進行因素的萃取，並以最大變異法轉軸方式來旋轉因素軸，因素個數的選擇則以特徵值(eigenvalue) 大於 1 者為判斷依據。接著，查看各因素轉軸後之因素負荷量 (factor loading)，選擇負荷量大於 0.399 的題項形成一個因素並刪掉落單的題目和只有兩題的向度，直到因素結構趨於穩定始停止，共計剩下 70 題，題號重新排列。特徵值大於 1 的因素共有 16 個，根據轉軸平方和負荷量總和，刪去小於 2 之因素，保留其餘因素 14 個。這些因素共解釋了 76.400% 的變異量。大部分的因素結構符合理論架構，只有因素一與因素二包含不同向度，但仔細觀察因素一與因素二，發現這些題項有其共通性，譬如因素一包含「社交、吸引、社群、分享」皆有「交流」的共通點，其中屬於「吸引」的題目是「線上閱讀的多媒體環境吸引我」強調了線上閱讀的特性，因此整個向度可以重新命名為「網路社群交流」。而因素二包含「認可、表現、競爭、推薦」，題目的共通性皆為「為了外在某一目的而閱讀」，目的包含為了「學業表現」或「為了在他人面前表現」，因此命名維持「表現」。

4.3. 信度分析與向度命名

因素架構穩定後，進行信度分析，以了解向度間的一致性，根據 Nunnally (1978) 建議 Cronbach's α 值之判準為大於 0.7 為佳，大於 0.6 為勉強接受，除了「好奇」向度之 Cronbach's α 為 .686 稍低之外，其餘因素皆達 .7 以上。本量表由網路特性與傳統閱讀向度構成，經過因素分析重新命名後，網路特性的向度共有 3 個向度，分別為社群交流、回饋與即時便利性；傳統閱讀共包含 11 個向度，分別為表現、推薦、滿足幻想、閱讀效能、順從、閱讀享受、逃避閱讀、認可、挑戰、反思、好奇。整體問卷的信度為 .957。

5. 討論與結論

本研究為一初探性研究，目的在發展一份具有信度與效度的線上閱讀動機量表(Motivation for Online Reading Scale, MORS)，希望未來能透過此量表了解使用者在線上閱讀的動機與行為。本量表在信度部分以內部一致性係數反映本測量工具具有內部同質性與一致穩定性，Cronbach's α 值皆有 .6 以上，甚至達到 .9；效度部分根據文獻與專家檢核，確立表面效度與內容效度，顯示題目具有邏輯性，測驗內容具有一定的廣度、涵蓋性與豐富性，也具有理論依據。此外，因素分析的結果，也顯示本問卷符合理論結構，具有因素效度。此一線上閱讀問卷共包

含 14 個向度：社群交流、回饋、即時便利性、表現、推薦、滿足幻想、閱讀效能、閱讀享受、逃避閱讀、認可、挑戰、反思、好奇，共 70 題。

謝誌

感謝國科會科教處對本研究的贊助，計畫編號為 NSC 97-2511-S-008-003-MY3

參考文獻

- 江靜之 (2003)。網際網路的衝擊。台北縣：韋伯文化國際出版。
- 宋曜廷、劉佩雲和簡馨瑩 (2003)。閱讀動機量表的修訂及相關因素研究。《測驗學刊》，51(1)，47-71。
- 李素足 (1998)。台中縣市國小中、高年級學童閱讀動機的探討。臺中師範學院國民教育研究所碩士論文，未出版。
- 洪蘭和曾志朗 (2001)。兒童閱讀的理念---認知神經心理學的觀點。《教育資料與研究》，38，1-4。
- 張春興 (1988)。知之歷程與教之歷程：認知心理學的發展及其在教育上的應用。《教育心理學報》，21，17-38。
- 陳俞霖 (2003)。網路同儕對 N 世代青少年的意義—認知感的追尋。嘉義縣：南華大學社會研究所。
- 黃馨儀 (2002)。國小學童閱讀動機量表之編製與相關研究。國立臺南師範學院國民教育研究所碩士論文。未出版。
- 蔡慧美 (2005)。國中生課外閱讀行為與經驗之研究。國立臺灣大學圖書資訊學研究所碩士學位論文，未出版。
- 賴怡君、張瑜芳、陳峰毅和劉旨峰 (2009)。青少年部落格分析：如果年輕。台灣師範大學：全球華人計算機教育應用大會(GCCCE 2009)。
- Baker, L., & Brown, A. L. (1984). Metacognitive skills and reading. In P. D. Pearson (Eds.), *Handbook of reading research* (pp. 353-394). New York: Longman.
- Guthrie, J. T., McGough, K., & Wigfield, A. (1994). *Measuring Reading Activity: An Inventory*. (Instructional Resource No. 4). Athens, GA: National Reading Research Center.
- Nunnally, J. (1978), *Psychometric Theory*, New York: McGraw-Hill.
- Schutte, N. S. & Malouff, J. M. (2007). Dimensions of reading motivation: Development of an adult reading motivation scale. *Reading Psychology*, 28(5), 469 - 489.
- Vygotsky, L. S. (1978). *Mind in society: the development of higher psychological process*. Cambridge: Harvard University Press.
- Watkins, M. W., & Coffey, D. Y. (2004). Reading motivation: Multidimensional and indeterminate. *Journal of Educational Psychology*, 96, 110-118.
- Wigfield, A. & Guthrie, J. T. (1995). *Dimensions of Children's Motivation for Reading: An Initial Study*. Athens, GA: National Reading Research Center.
- Wigfield, A. & Guthrie, J. T. (1997). *Reading engagement: Motivation readers through integrated instruction*. Newark, DE: International Reading Association.
- Wigfield, A., Guthrie, J. T., & McGough, K. (1996). *A questionnaire measure of children's motivations for reading*. (Instructional Resource No. 22). Athens, GA: National Reading Research Center.

國小五年級數學領域概數與估算單元數位個別指導模式之研發

The Development of Digital Learning Content for Individual Instruction — using the “round number and estimation” unit in 5th grade as an example

許天維

國立台中教育大學教育測驗統計研究所教授

郵件信箱：sheu@mail.ntcu.edu.tw

郭伯臣

國立台中教育大學教育測驗統計研究所教授

郵件信箱：kbc@mail.ntcu.edu.tw

劉育隆

亞洲大學資訊工程學系博士班

郵件信箱：bms094112@ms3.ntcu.edu.tw

【摘要】本研究主旨在研發數位個別指導補救教材，且配合「以知識結構為基礎」之電腦適性診斷測驗進行研究，並將受試者分為教師補救與電腦補救，觀察受試者進行補救教學後是否因而有所進步，再探討兩種不同補救模式的成效，本研究結果如下：

- 1.無論是經由教師補救或者電腦補救後，學生的平均分數皆進步4分以上，據結果顯示數位個別指導補救教材運用於補救教學有一定之成效。
- 2.研究結果顯示教師補救教學低分群與電腦補救教學對中分群、低分群的學生，在後測時進步的程度都有到達顯著。

【關鍵詞】電腦適性測驗、補救教學、概數與估算

***Abstract:** In this research, a hybrid method of computerized adaptive test and individual remedial instruction is proposed. The computerized adaptive testing algorithm is based on students' item structure and the individual remedial instruction is designed by experts' knowledge structure and students' item structure. Computerized adaptive test is first administrated for diagnosing individual student's learning profile. Based on this profile, the adaptive individual remedial instruction process is constructed and administrated. The difference of pre-test and post-test shows that the proposed method can help students to improve their learning situations.*

Keywords: Computerized Adaptive Test, Individual Remedial Instruction, Round Number and Estimation

1.前言

適性化、個別化的教學評量已是近年來測驗的趨勢，在資訊化的教育環境下，結合電腦的適性診斷測驗「因材施教」，更能節省時間、準確的預估受試者的學習狀態，提供教學者瞭解受試者的學習狀態、精確的補救訊息，讓教學者進行有效率的補救教學，達到「因材施教」的目標。本研究「概數與估算」單元為例，開發以「知識結構為基礎」的數位個別指導教材，在學生進行電腦化適性測驗後，根據個別診斷報告提供數位個別指導，嘗試精確的由下位概念至上位概念進行補救教學，提供即時回饋，有效的幫助學生進行概念澄清。最後並探討不同的補救教學模式的成效。

2. 文獻探討

本研究是以研發「以教師為基礎」之數位個別指導教材(Teacher- Based adaptive Remedial Instruction, TBRI)，編制以「概數與估算」單元為例之教材，配合「以知識結構為基礎」之電腦適性診斷測驗(knowledge structure based adaptive testing, KSAT)進行研究，探討 TBRI 對於補救教學的成效，因此本章節就以 KSAT 與知識結構進行文獻探討。

2.1 以知識結構為基礎之電腦化適性診斷測驗 (KSAT)

學生在進行「以知識結構為基礎之適性測驗」後立即能提供成績回饋，達到「因材施教」的效果，並有相關研究指出電腦化適性測驗確實可以節省施測題數、時間，且有不錯的精準度（蔡昆穎，2004；許志毅，2004；黃碧雲，2005；趙琬津，2006）。KSAT 雖然可以在短時間內對學生的學習情況進行診斷，了解學生的知識結構是否有不足之處，但在獲得學童的學習診斷並非教育的最終目的，最終還是需要學會完整的教材內容，因此測驗後之補救教學實有必要。

2.2 知識結構(knowledge structure)

知識結構可分為「專家知識結構」、「學生知識結構」與「補救教學知識結構」。

「專家知識結構」是由多位學科專家根據教學理論與實地教學經驗，分析在測驗範圍內的知識概念，再根據學生的學習歷程、概念發展順序以及概念之間的上下位關係整理而成的結構關係（郭伯臣，2003）。

「學生知識結構」或「學生試題結構」是運用專家知識結構編製而成的紙筆測驗，進行施測後，根據得到的資料，以「順序理論」(ordering theory, OT) (Airasian & Bart, 1973)分析估計而得。學者郭伯臣、謝友振等、張峻豪、蔡坤穎（2005）分析估計三種試題結構的方法，研究結果指出以「順序理論」的適性測驗演算法在「節省試題」和「預測精準度」兩方面都有最佳的表現，優於「試題關聯結構法」(item relationship structure theory, IRS) (Takeya, 1991) 與 Diagnosys (Appleby, Samuels, Treasure Jones, 1997)，因此本研究採用「順序理論」的適性測驗演算法。

3. 研究方法

將就整個研究對象、研究工具、實驗設計、研究流程和資料處理與分析加以討論，內容說明如下：

3.1 研究對象

電腦適性測驗採方便抽樣，對象為九十七學年度已上完「概數與估算」單元的國小五年級學生。因考慮研究者時間及行政上的支援等因素，因此選擇台中縣某國小五年級 3 個班級共計 71 名學生為研究對象。

3.2 研究工具

每位學生在接受前測、後測後可以得到一份個別診斷報告，教師會針對該學生的錯誤概念進行補救教學，或直接於觀看動畫進行適性補救教學，如圖像 1 所示。



圖像 1 KSAT 施測畫面、個別診斷報告及適性補救教學動畫

3.3 實驗設計與流程

本研究採準實驗設計(quasi-experimental design)，總共實驗 3 個班，對照組為 3 個班級隨機抽出 12 位學生，共 36 位，其餘學生為實驗組。實驗組的學生採用電腦補救教學，對照組則是回到班級進行團班補救教學，實驗流程圖，如表 1 所示。

表 1 實驗流程表

時間	實驗組	對照組
10 分鐘	系統使用說明	
20~30 分鐘	KSAT 前測	
40 分鐘	數位個別指導(電腦補救教學)	團班上課
20~30 分鐘	KSAT 後測	

4. 研究結果

4.1 「以電腦為基礎之適性補救教學」之成效

表 2 為前、後測平均分數比較表，其中實驗組是電腦補救教學，共 35 人；對照組是教師補救教學，共 71 人。而實驗組與對照組中，個別再細分高分組、中分組、低分組。表 3、表 4 分別為教師補救教學與電腦補救教學的前、後測成績。

表 2 前、後測平均分數比較表

組別	個數	平均分數			顯著性
		前測	後測	進步	
電腦補救(實驗組)	35	88.0000	92.7714	4.7714	.000
教師補救(對照組)	36	86.6667	90.6667	4	.004

由表 2 發現，不論是實驗組或是對照組，再經過補救教學後，學生的平均分數皆進步 4 分以上。

表 3 電腦補救教學前、後測成績

組別		高分群後-前測	中分群後-前測	低分群後-前測
平均成績	前測	96.75	90.5	75.7273
	後測	97	95.75	84.9091
	進步分數	0.25	5.75	9.1818

表 4 教師補救教學前、後測成績

組別		高分群後-前測	中分群後-前測	低分群後-前測
平均成績	前測	94.25	87.3333	78.4167
	後測	94.75	89.75	87.5
	進步分數	0.5	2.4167	9.0833

由表 2 發現，不論是實驗組或是對照組，再經過補救教學後，學生的平均分數皆有進步。由表 3、表 4 發現，將前、後測成績相比較，不論是電腦補救教學或教師補救教學，兩者低分組的前、後測成績有顯著性差異，且而電腦補救教學中分組的後測成績明顯優於前測成績。

4.2 兩種不同適性補救教學模式之成效

由表 3、表 4 發現，不論是電腦補救教學或教師補救教學對於高分組與低分組成效差不多，而電腦補救教學對於中分組的補救成效較優於教師補救教學。

5. 結論與建議

5.1 結論

(一)、依據使用者的需求觀點，「以電腦為基礎」之數位個別指導教材具有可用價值。

(二)、不論是電腦補救教學或教師補救教學，經過相同教材的補救教學後，學生的平均分數皆進步 4 分以上，顯示教材結合補救教學有一定之成效。值得注意的是本研究研發之數位指導教材適用於實驗中所有學生，對低、中、高成就的學生皆有進步的效果。

本研究的數位個別指導教材配合以教師為基礎的適性補救教學，具有一定之效果，輔以數位多媒體教材，更能吸引學生的注意力，提高補救教學成效，讓中低分組的學生皆有明顯的進步。

參考文獻

- 郭伯臣（2003）。國小數學科電腦化適性診斷測驗(I)。《行政院國家科學委員會專題研究計畫報告 (NSC 91-2520-S-142-001)》。
- 許志毅（2004）。國小數學領域電腦化適性診斷測驗及補救教學系統之內容開發及試用-以「扇形」單元為例。《國立台中師範學院教育測驗統計研究所教學碩士論文》。
- 黃碧雲（2005）。以能力指標結構為基礎的電腦適性測驗編製及動畫補救教學之應用—以國小數學領域四年級能力指標。《國立台中師範學院數學教育學系碩士班碩士論文》。
- 趙琬津（2006）。數位個別指導教材與模式之開發-以三角形單元為例。《台中教育大學數學教育學系在職進修教學碩士論文》。
- 蔡昆穎（2003）。國小數學領域電腦化適性診斷測驗及補救教學系統之內容開發及試用—以「擴分、約分」單元為例。《台中師範學院測驗統計研究所碩士論文》。
- Airasian, P. W., & Bart, W. M. (1973). Ordering theory: a new and useful measurement model. *Educational Technology*, May, 56-60.
- Appleby, J., Samules, P., & Treasure-Jones, T. (1997). Diagnosys a knowledge-based diagnostic test of basic mathematical skills. *Computer Education*, Vol.28, No.2, pp.113-131.
- Takeya, (1991). New item structure theorem. Tokyo: Waseda University.

改良式選擇題題型之作文能力測驗方法研究

梁惠玲

臺北市立教育大學附設實驗國民小學，臺北，臺灣

邮件信箱：t8906@estmue.tp.edu.tw

孫劍秋

國立臺北教育大學語文與創作學系，臺北，臺灣

邮件信箱：sun0761@tea.ntue.edu.tw

吳偉賢

國立臺北教育大學資訊科學系，臺北，臺灣

邮件信箱：wsu@tea.ntue.edu.tw

楊志強

國立臺北教育大學教育學系，臺北，臺灣

邮件信箱：cyang@tea.ntue.edu.tw

【摘要】本研究提出一個改良式選擇題題型測驗及其計分方式，每一題之不同選項配予不同分數，期能測出學童對同一題目更精確的程度差異，進而探討其與作文能力之間的關聯。研究結果顯示本研究發展之測驗與作文成績的相關係數為 0.533，相較於國語成績與作文成績的相關係數為高。

【關鍵詞】 改良式選擇題題型、選擇題測驗、語文能力、寫作能力

1.前言

傳統作文評定的弊端為缺乏可靠有效的測量工具和客觀的評斷標準（朱作人，1991），而選擇題題型一直被嘗試用來改進人工閱卷的缺失，尤其是大規模的檢測，選擇題題型可利用電腦進行判讀，節省人力、物力並可降低人為疏失的機率。在現行的選擇題測驗中，即使獲得同樣的分數，仍無法測出學生在特定情境下，真實的語文運用能力之差異。例如，春風又「綠」江南岸與春風又「到」江南岸，這兩個字的用法，並無絕對的對與錯，但有文字情境運用的差別。因此，本研究，試著發展一種不同於以往的，改良式的新的測驗題型，選擇題中的四個答案選項，各依其與題幹的適切度而得不等的分數，而能更真切反映學生之語文能力與作文能力之間的關係。

2.文獻探討

作文評量現況依據國中基測評量標準（教育部「95 年國民中學學生寫作測驗試辦實施方案」），以六級分距，分別從立意取材、結構組織、遣詞造句及錯別字、格式與標點符號等四大面向評定作文成績。

3.評量工具編製

本研究取立意取材、結構組織與遣詞造句為三個面向，作為評量的架構。由這三個面向來觀照學童的作文能力。本測驗共計 30 題選擇題，評量工具編製（歐滄和，2002）、（陳英豪與吳裕益，1998）在每個選項均呈現與題幹不同的適切程度，依據學生選答項目，判斷學生對

於題幹描述的情境其理解程度，期能具有鑑別學生的修辭、句型、語彙、推論、文意理解、掌握大意主旨、辨別文字正確形音義的功能。此改良式選擇題，每題題目各設計出四個答案選項，每一個答案選項依適切程度而獲得 4~1 不等的分數，依所得分數，判別學生不同的語文程度。

4.測驗實施

本研究以台北市國小五年級學童為研究對象，進行預試及正式施測，正式施測時受測者為兩校五班國小五年級學生，其中男生 80 人、女生 101 人。對第一所學校之同樣本實施作文實測，惟實施當時因校內活動，僅 72 人參與作文實測。

5.資料分析與結果

本研究之配分方式有別於傳統單選題型，因此尚無可直接套用之檢測分析工具。雖然如此，本研究仍試圖以現有檢測方式來檢驗（楊志強，2004），檢驗結果均具信效度及鑑別度。

國語學期總成績是由整學期平時和段考多種成績平均所得到的結果，具有極高的效度，因此本研究以國語總成績為效標，和改良式選擇題題型測驗總得分算出其相關係數，以考驗本評量工具的效度，其結果顯示改良式選擇題題型測驗與國語成績、作文成績、立意取材、結構組織、遣詞造句皆有顯著正相關，也就是說，本評量工具的設計具有良好效標關聯效度。研究結果同時顯示改良式選擇題題型測驗與作文實測有顯著正相關。本研究測驗總分與學生在校國語成績相較，在立意取材（0.477比0.308）、組織結構（0.491比0.267）、遣詞造句（0.483比0.253）及作文成績（0.533比0.327）各方面均有更高的關聯性。

6. 結論與建議

改良式選擇題題型測驗更真確表現學童的作文能力與程度，且比一般的國語科成績與立意取材、結構組織、遣詞造句這三個向度的作文能力更具相關性。本研究之改良性選擇題題型，語詞的認知具爭議性，因此，影響配分的認定與解釋，本研究之改良性選擇題給分，並非沒有再討論的空間。

將本研究之改良式選擇題題型，普遍應用於測驗上，勢必引領教師在教學或學生在學習上，更注意明辨語詞的意義與運用。不但能增進學生語文程度、精準語詞運用能力，而且也能帶動教師精進教學，提升語文教學內涵。且本研究之改良式選擇題題型測驗，較傳統測驗方式能避免只有唯一答案之思考模式，期待這樣的影響具有正向的影響。

本研究之改良式選擇題型因配分方式有別於傳統單選題，因此使用現有的信度、鑑別度評鑑方式並不能充分反應本測驗，本研究雖以其他之數據、圖表作為依據，利用現有之方式探討其信度及鑑別度，惟並非經過嚴謹之學術研究驗證，此為發展推廣本研究成果之重要課題。

誌謝

本研究接受國科會補助（NSC98-2511-S-152 -018 -MY2），謹致謝忱。

參考文獻

- 朱作人主編（1991）。《語文測驗原理與實施法》。上海：上海教育出版社。
 陳英豪、吳裕益（1998）。《測驗與評量》。高雄市：復文圖書出版社。
 歐滄和（2002）。《教育測驗與評量》。台北市：心理出版社股份有限公司。
 楊志強（2004）。測驗品質考驗與 TestGraf 98 的應用。《教師專業成長與實踐智慧》，頁 93-104。

建置攝影課程作品線上評量系統

The Development of a Web-Based Assessment System for Photography

吳振宏¹、蘇彥寧²

國立臺南大學教育學系科技發展與傳播碩士班

郵件信箱：{m09839012, m09739012}@stumail.nutn.edu.tw

歐陽閻³

國立臺南大學教育學系副教授

郵件信箱：ouyang@mail.nutn.edu.tw

【摘要】本研究旨在建置一攝影課程作品線上評量系統，俾輔助教師快速便利評量學生作品。本研究以吳振宏、蘇彥寧、歐陽閻（2009）發展之「攝影課程作品評量規準」作為評量準則，並依此開發「攝影課程作品線上評量系統」。最後透過形成性評鑑，採專家意見檢核本系統並進行修正。結果顯示專家對於線上評量系統的介面設計、系統功能及整體滿意度等三大面向多持正向意見。具體而言，本系統具備「易管理」、「易瀏覽」、「使用便利」、「評量快速」、「高支援性」等效益，可協助教師快速便利評量學生作品。未來期透過實務驗證，以評估本系統應用於實際課程評量之成效。

【关键词】線上評量、攝影課程、攝影作品評量、形成性評鑑

Abstract: The aim of this study was to develop a Web-Base Assessment System for Photography, formative evaluation was used in which as the method, of using Photography Work Assessment And gauge (Wu, 2009) in support for instructor save time of evaluate the photography-related courses opus for students. The main result showed that most specialty have optimistic view and feel satisfying about Web-Based Assessment System of interface design, system function and whole satisfying. In other words, the characters of this system have to easy manage, easy review, easy conduct, quickly evaluate and high supportive, which support for instructor save time of evaluate the photography-related courses opus for students. In the future, we will use examine to assess the system.

Keywords: Online Assessment, Photography Course, The Assessment of Photography, Formative Evaluation

1.緒論

近年來攝影器材與過去相比，現今攝影器材售價低廉且逐漸普及。此一發展趨勢使得愈來愈多的人開始接觸與喜歡攝影，且有將「攝影」視為社會上文化休閒活動之趨勢（黃嘉勝，1999）。在攝影相關課程逐漸受到重視與推廣的同時，各攝影相關課程教師面對數量龐大的學生數位作品，如何快速瀏覽與準確、方便的進行評量，已成為攝影相關課程教師教學上所面臨的困境。因此，若能將線上評量的概念導入課程中，並建置一套適合於攝影相關課程的線上評量系統，將能使教師更方便與有效率的評量學生的作品。

2.線上評量系統之規劃及實作

本研究所提出之「攝影課程作品線上評量系統」，依課程角色之不同，分為教師(授課)、教師(評量)及學生三類，並另提供系統管理員身份，作為平台管理時使用，服務架構如圖 1 所示。

本系統為 Web-Based 架構，修課學生及教師可透過網路連結至本研究所開發之「攝影作品線上評量系統」。教師登入系統後可依據需求，使用「作業管理」、「繳交狀況」、「作品評量」及「匯出成績」等功能，編輯與控管作業的各類資訊。

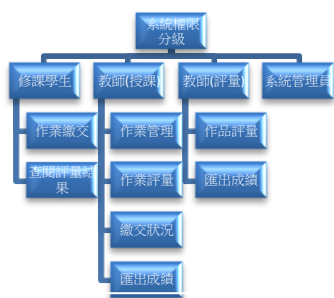


圖 1 服務架構

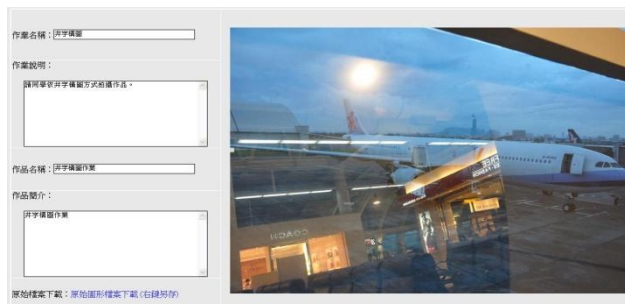


圖 2 學生作品評量上半部頁面

門牌	1. 能準確掌握門牌特徵，能下決定性的一一。	1. 無法掌握門牌特徵，能下決定性的一一。
窗戶	2. 能準確掌握窗戶特徵，能下決定性的一一。	2. 無法掌握窗戶特徵，能下決定性的一一。
牆壁	3. 能準確掌握牆壁特徵，能下決定性的一一。	3. 無法掌握牆壁特徵，能下決定性的一一。
地板	4. 能準確掌握地板特徵，能下決定性的一一。	4. 無法掌握地板特徵，能下決定性的一一。
色彩	5. 能準確掌握色彩特徵，能下決定性的一一。	5. 無法掌握色彩特徵，能下決定性的一一。

教師可透過「作品評量」的功能，進入評量學生作品的頁面。該頁面的上半部會顯示學生的作品名稱、簡介及圖檔，並於左下方提供原始圖檔的下載，以提供特殊圖檔格式無法產生縮圖的解決方案。頁面下半部為評量規準之量表，該量表參考吳振宏等(2009)所提出之攝影作品評量規準，提供教師評量學生作品時之參考。如圖2、

3 所示。圖 3 學生作品評量下半部頁面

3.線上評量系統之形成性評鑑結果分析

本研究為探究此一線上評量系統是否具有預期功能，遂進行形成性評鑑。故研究者分別邀請 10 位擁有攝影教學專業領域之實務與（或）研究經驗之專家學者作為審查委員，進行線上審查。問卷針對介面設計、系統功能、整體滿意度三大面向調查結果顯示，每一面向之正向滿意度合計皆達 80% 以上。此一結果顯示本研究所建置之系統受到專家的高度肯定。

4.結論

本研究建置之系統可協助攝影課程教師快速便利評量學生作品，並具備下述之相關效益：

- 一、易管理：教師可透過「作業管理」功能，進行作業指派、編輯作業名稱與說明，以及調整各項作品評量規準的權重值，並對學生繳交作業的權限進行控管。
- 二、易瀏覽：教師可選擇「單一作業」或「個別學生」進行作品瀏覽，系統即會呈現作品名稱、作品縮圖及分數(已評量)，方便教師瀏覽學生作品。
- 三、使用便利：可透過現行的瀏覽器連結系統，方便教師、學生進行各項活動。
- 四、評量快速：供教師於評量學生作品時可以交互參照作品與規準，並直接於同一頁面評量。使評量過程得以簡化及更加快速。
- 五、高支援性：系統對於學生上傳作品的檔案格式並無限制，且系統提供三種常見檔案格式(JPG、JPEG、PNG)的自動精細縮圖服務，使學生上傳作品時無須另外轉檔縮圖。

參考文獻

- 吳振宏、蘇彥寧、歐陽閻 (2009)。建構攝影課程作品評量規準之初探。載於國立彰化師範大學舉辦之「TANET 2009 台灣網際網路學術研討會論文集」(頁 I1-I6)，彰化市。
- 黃嘉勝 (1999)。攝影作品鑑賞教學的設計與應用之研究。載於崑山技術學院舉辦之「第十四屆全國技術及職業教育研討會論文集」(頁 125-134)，台南縣。
- DC View (2009.07.17) 取自 <http://www.dcvview.com/>。

Reducing the Impact of Inappropriate Items on Reviewable CAT

Yung-Chin Yen, Rong-Guey Ho, Li-Ju Chen, and Wen-Wei, Liao

Graduate Institute of Information and Computer Education, National Taiwan Normal University

scorpio@ice.ntnu.edu.tw, hrg@ntnu.edu.tw, ljchen@ice.ntnu.edu.tw, abard@ice.ntnu.edu.tw

Abstract: The underlying hypothesis of reviewable CAT was that after rereading or rethinking an item, the examinees might correct the careless mistake they made. However, changing the answer of one item in CAT might cause the following items no longer appropriate for estimating the examinee's ability. These inappropriate items in a reviewable CAT may introduce bias in ability estimation and decrease precision. This study attempted to evaluate the performance of four-parameter logistical (4PL) model by comparing it with three-parameter logistical (3PL) model and utilizing it to reduce the impact of inappropriate items on reviewable CAT.

Keywords: IRT, reviewable CAT, upper asymptote parameter, 4PL IRT model, rearrangement procedure

1. Introduction

The underlying hypothesis of reviewable CAT was that after rereading or rethinking an item, the examinees might correct the careless mistake they made. This hypothesis afterwards led to the fact that even high-ability students may on occasion miss items that they should have answered correctly. However, changing the answer of one item in CAT might cause the following items no longer appropriate for estimating the examinee's ability. These inappropriate items in a reviewable CAT may introduce bias in ability estimation and decrease precision. The same situation was also seen in traditional CAT. In virtue of the underlying characteristics of the traditional IRT model and the item selection method, an examinee's ability would be considerably underestimated if he/she missed early items. To cope with the underestimation problem, Barton and Lord (1981) proposed the four-parameter logistical (4PL) IRT model allowing a high-ability student to miss an easy item without having his ability drastically lowered. This study attempted to compare the performance of 4PL and 3PL (three-parameter logistical) IRT model and tried to implement 4PL model as solution for inappropriate items confronted in reviewable CAT.

2. Literature Review

Item response theory was a family of mathematical descriptions describing what happens when an examinee meets an item. According to the number of item parameter, IRT model can be generally classified into three widely used categories: one-parameter logistic (1PL) model, 2PL model, and 3PL model.

2.1. One-, two-, and three-parameter logistical IRT model

In 1PL model, the probability that an examinee with ability θ can answer an item with difficulty b correctly can be expressed as $P_{1PL}(\theta) = 1/(1 + \exp[(\theta - b)])$ where D is a scaling factor whose value is 1.702. The mathematical form of the 2PL model could be written as $P_{2PL}(\theta) = 1/(1 + \exp[-Da(\theta - b)])$ while the new parameter a in 2PL model is called the discrimination parameter which allowing an item to discriminate differently among the examinees. The probability of $P_{2PL}(\theta)$ ranges from zero to one as θ goes from $-\infty$ to ∞ .

On a multiple-choice test, however, the probability of choosing the correct answer does not approach zero for low-ability students (Barton & Lord, 1981). Even an examinee who knew nothing still had a one-out-of-four chance to choose the correct answer in a multiple choice test with four options (Yen, Ho, Chen, Chou, & Chen, in press). Birnbaum (1968) introduced a guessing parameter c to handle the situation in which examinees either guess totally randomly or answer on the basis of their knowledge. The resulting 3PL model is $P_{3PL}(\theta) = c + (1 - c)P_{2PL}(\theta)$.

2.2. Four-parameter logistical IRT model

To address examinees' careless mistakes in CAT, Barton and Lord (1981) proposed the 4PL IRT model which introduced an upper asymptote, expressed by the Greek letter delta (δ), into the 3PL model:

$P_{4PL}(\theta) = c + (\delta - c)P_{2PL}(\theta)$. While $P_{2PL}(\theta)$ ranges from zero to one, $P_{4PL}(\theta)$ ranges from the lower asymptote, c , to the upper asymptote parameter, δ , for item-specific "carelessness". To evaluate the effect of the upper asymptote on ability estimation, Rulison and Loken (2009) conducted two CAT simulation experiments to compare 3PL model with 4PL model in regard to estimation bias and root mean square error (RMSE) for high-ability students with a poor start (an examinee missed the first two items). According to Rulison and Loken's study, using 4PL model ($\delta = 0.98$) can lower bias and RMSE for high-ability student with a poor start. In other words, 4PL IRT model proposed examinees an opportunity to recover from inappropriate responses in CAT.

2.3. Traditional Solutions for Reviewable CAT

The term "item review" in testing contexts referred to administrative rules that allowed examinees to change their responses to previously answered items. Prior research has shown that the examinees tend to increase their test scores when they are allowed to revise their answers. In a CAT, however, it was often assumed that examinees should not be allowed to review previous items and change answer. The reasons for this range from the bias in ability estimation, the

potential to obtain artificially inflated scores, reduced testing efficiency, item dependence, to extra complexity in the item selection algorithm.

Vispoel, Hendrickson, and Bleiler (2000) proposed the limiting answer review and change procedure that allowed reviewing and changing within successive m-item blocks. In this procedure, examinees were only allowed to review and change answers within the recent block. If an examinee was answering the items in block j , he/she was not allowed to review the items in the previous blocks. According to their study, this procedure can overcome the examinees' cheating strategies in reviewable CAT without diminishing estimation precision.

Another solution referred to dropping inappropriate items in reviewable CAT. Papanastasiou (2002) proposed the rearrangement procedure which rearranged and skipped certain items to better estimate the examinees' abilities for a reviewable CAT. Three types of answer changing caused the rearrangement procedure in ability estimation. Type 1 change involves changing answers from incorrect to incorrect which do not need re-estimating ability. The second type involves changing answers from incorrect to correct and it would result in item skipping in the rearrangement procedure. The third type is to change answers from correct to incorrect and it would also result in item skipping in the rearrangement procedure.

The underlying hypothesis of reviewable CAT led to the fact that high-ability students may on occasion miss items that they should have answered correctly. However, almost all previous reviewable study was conducted based on traditional CAT assumed that a high-ability student should answer an easy question with probability approaching one. The underlying hypothesis of reviewable CAT was consistent with the principle of 4PL model. Besides, the 4PL model may propose examinee an opportunity to recover from the inappropriate responses which introduced by reviewing and changing answer in reviewable CAT. In the present study, therefore, the effect of 4PL model and the rearrangement review solution on reducing estimation bias was investigated.

3. Method

Three experiments will be conducted in this study. The first two experiments, a simulation (experiment 1) and an empirical one (experiment 2), focused on a study of evaluating the effect of upper asymptote on CAT by comparing the measurement precision and efficiency of 3PL and 4PL IRT model under both simulation and empirical conditions; the third one focused on the study of comparing the performance of 4PL and rearrangement on reducing the estimation bias introduced by inappropriate items on reviewable CAT.

In experiment 1, 200 simulees sampled from all 13000 simulated examinees will take the following four different versions of CAT: P3CAT (3PL CAT with poor start), P4CAT (4PL CAT with poor start), N3CAT (normally administered 3PL CAT), and N4CAT (normally administered 4PL CAT). The estimation precision and efficiency of these four CATs will be compared to investigate the performance of 3PL and 4PL on estimation precision and efficiency. Experiment 2 tries to investigate the same questions with empirical examinees and item bank.

In experiment 3, same examinees as experiment 1 will be randomly assigned into following four different CATs: R3CAT (3PL reviewable CAT), R4CAT (4PL reviewable CAT), RR3CAT (3PL-based reviewable CAT implementing rearrangement procedure), and RR4CAT (4PL-based reviewable CAT implementing rearrangement procedure). This purpose of this experiment is to compare the performance of 4PL IRT and rearrangement procedure on reducing the inappropriate-items effects introduced by reviewable CAT.

4. Conclusion

All three experiments are under development and the final result will be given later. If the experiments results support our hypothesis, the 4PL model should attract more attentions and be applied not only in reviewable CAT, but also in other testing contexts.

References

- Barton, M. A., & Lord, F. M. (1981). *An Upper Asymptote for the Three-Parameter Logistic Item-Response Model*. Princeton, NJ: Educational Testing Services.
- Birnbaum, A. (1968). Some latent traits model and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.
- Papanastasiou, E. C. (2002, April). *A 'rearrangement procedure' for scoring adaptive tests with review options*. Paper presented at the the National Council of Measurement in Education, New Orleans, LA.
- Rulison, K., & Loken, E. (2009). I've fallen and I can't get up: can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement*, 33(2), 83.
- Vispoel, W. P., Hendrickson, A., & Bleiler, T. (2000). Limiting answer review and change on computerized adaptive vocabulary tests: Psychometric and attitudinal results. *Journal of Educational Measurement*, 37(1), 21–38.
- Yen, Y. C., Ho, R. C., Chen, L. J., Chou, K. Y., & Chen, Y. L. (in press). Development and evaluation of a confidence-weighting computerized adaptive testing. *Educational Technology & Society*.

基於本體論的形成性評量應用於戶外無所不在唐詩教學之成效研究

A Study of Ontology-based Formative Assessment for Ubiquitous Tang Poetry Learning in Outdoors

時文中、曾憲雄¹

亞洲大學 資訊多媒體應用學系
{wjshih, sstseng}@asia.edu.tw

【摘要】傳統的國小唐詩教學有兩個困難。首先，強調解釋與背誦的教學方式僅能達到認知與技能面向的目標，不容易讓國小學童體會教材中的意境與達成情意面向的目標。其次，由於小學生的語文表達能力有限，老師不容易得知學生對於教材意境的理解程度，以及學生表達用詞的意向。因此，本研究嘗試在戶外進行無所不在唐詩教學，並提出「基於本體論的形成性評量」來促進唐詩教學的成效。實驗結果顯示使用本教學法的學童在學習成效有顯著提升。問卷調查結果也顯示此系統能協助教師即時得知學生的學習狀況。

【關鍵詞】數位學習、無所不在學習、唐詩教學、本體論、形成性評量

Abstract: Two drawbacks exist in conventional Tang Poetry instruction. First, traditional instruction using explanation and recitation may achieve the cognitive and technical objectives. But the affective objective may not be fulfilled. Second, restricted by the immature ability to express their feelings, elementary-school children's intention can not be easily understood by the teachers. This work proposes Ontology-based formative assessment and applies this technique to ubiquitous Tang poetry instruction in outdoors. Experimental results show that the learning performance of students are improved using this approach. Also, Surveys show that this system can help teachers understand learners' status about the poems.

Keywords: e-Learning, Ubiquitous learning, Tang poetry instruction, Ontology, Formative assessment

1.簡介

布魯姆等教育學家認為教育目標包括：認知、技能與情意三個面向，其中情意面向涉及情感陶冶與人格形成，尤其重要[1]。然而，傳統的國小教學有兩個困難。首先，強調解釋與背誦的教學方式僅能達到認知與技能面向的目標，不容易讓國小學童體會教材中的意境與達成情意面向的目標。其次，由於小學生的語文表達能力有限，老師不容易得知學生對於教材意境的理解程度，以及學生表達用詞的意向。因此，針對國小教學，如何提供適合實現情意目標的學習環境，以及如何探尋小學生遣詞用字的表達意向，成為急待克服的問題。

上述教學場景的關鍵之一是系統與學習者的人機互動。如果缺乏一個友善的人機互動介面工具，學童的學習狀態與意向將不容易得知，進而造成分組學習的效率不佳，需要仰賴老師投入教學活動，增加老師負擔。因此，本論文針對此關鍵需求，特別研製一個「互動式照片標籤法」來協助學童輕鬆地表達對於學習的感受。當學童可以輕鬆透過此工具與系統互動溝通，並表達學習感受時，老師就可以有更多餘裕規劃多元化教學活動，提升學習成效。

本研究之主要貢獻在於所研發一個「基於本體論的形成性評量」，可透過簡單的拍照與文字標籤，讓學生輕鬆地表達對於唐詩的情境感受。同時，老師也可以藉由這項回饋資訊了解學生的學習意向，進而提升學習成效。因此，本論文不但研發了人機互動的介面與工具，也促進學習科技與工具創新研發之發展目標。

2.系統實作與實驗結果

本論文實作一個「互動式照片標籤法」學習意向探尋介面工具，應用在環境感知的唐詩

¹ 通訊作者

教學環境。為了加速開發過程，我們採取「系統整合」的方式，利用開放程式碼的現有程式來實作系統中的各個模組。系統首頁如圖1所示，使用者可輸入關鍵字標籤或上傳照片：

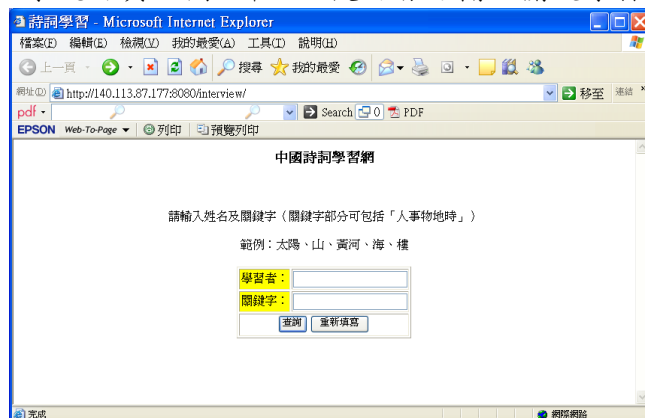


Figure 1. 「互動式照片標籤法」學習意向探尋系統之首頁

我們先實施前測以確定兩組學童在唐詩學習成就無顯著差異。表1為後測的t檢定結果。 $|t| = 1.98 > t_{\alpha}(9) = 1.833$ ，表示A與B組的後測平均分數存在顯著差異。也就是說，B組在實施無所不在的唐詩教學後，學習成效顯著優於A組。

表 1. 後測的 t 檢定結果($\alpha=0.05$)

	人數	平均分數	標準差	t
A (控制組)	10	83.40	16.08	
B (實驗組)	10	91.70	5.83	
A - B	10	-8.30	13.23	-1.98

3. 結論

本研究研製的戶外無所不在學習系統利用行動裝置的照相、簡訊與通訊功能，讓學童能輕易地使用標籤註解來表達內心的概念。教師也可以透過比對學生與教材的本體論差異來評量學生的迷思概念，進而即時提供適當的教學指引。從前述的實驗結果發現，透過照相標籤工具的互動，系統可以更準確預測學童所要表達的概念。結果，老師可以適時提供適當的提示來解決學童的問題。因此，本研究發現戶外實境融入唐詩教學可以加強學生的學習成效。

誌謝

本研究承蒙國科會（計畫編號 NSC-97-2511-S-468-003, NSC-98-2511-S-468-002, NSC97-2511-S-009-001-MY3, NSC98-2511-S-468-004-MY3）補助經費支持。

參考文獻

- [1] G. J. Hwang, "Characters, Characteristics and Strategies of Ubiquitous Learning", IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC 2006), June 5-7, 2006, Taichung, Taiwan.