Is Teacher Web-based Portfolio Assessment Reliable or Valid in Computer Course?

Chi-Cheng Chang
Distinguished Professor
National Taiwan Normal University
Taipei, Taiwan
E-mail: samchang@ntnu.edu.tw

Kuo-Hung Tseng
Chair Professor
Meiho University
Pingtung, Taiwan
E-mail: gohome8515@gmail.com

Conference Key Areas: Information and communication technologies
Keywords: Web portfolio, e-Portfolio, Portfolio assessment, Reliability, Validity

1 INTRODUCTION

In discussing reliability issues portfolio assessment, Chang [1] stated that rubrics are not easy to manipulate when utilizing portfolio assessment. Lu [2] further pointed out there are several factors making the reliability dubious, including inconsistent scoring rubrics, ambiguous comments and limited capacity of student self- and peer-assessment. In validity research, Chang [3] indicated that with little qualitative and quantitative research available, portfolio studies on reliability and validity primarily rely on the self-narratives from students and instructors to investigate teaching and learning effectiveness. As reviewing previous investigations of reliability and validity, Derham and Diperna [4] stated that digital portfolios yet remain unexplored despite considerable amount of evidence available for paper-based formats. However, due to similar ways used to examine these two types of assessments, reliability and validity evidence based on paper assessments is still valuable to its counterpart [5]. It is evident the establishment of rubrics will increase reliability and validity of digital portfolio assessment; however, the level it achieves and whether or not it is sufficient still remain unknown. It is a matter of great urgency developing assessment rubrics and verifying their reliability or validity.

In order to explore a single dimension deeply, this study focused on the issues of teacher assessment. With above implications in mind, the purpose of this study was to investigate reliability and validity of teacher assessment under the web-based portfolio environment (or the Web-based teacher portfolio assessment) through a teaching experiment. Research questions are as follows: 1)Are the portfolio assessment results consistent among raters (teachers)? 2)Are the portfolio assessment results consistent within individual rater (teacher)? 3)Is each rubric aspect appropriate for examining learning achievements, i.e. are the portfolio scores consistent with the end-of-course examination scores?

2 METHOD

2.1 Participants

The participants were 79 eleventh-graders in the “Computer Application” course, among which the portfolios developed by 72 participants were completed and suitable for the statistical analysis. The participants, with basic computer skills including using the Internet, were taught 2 units of the computer course that addressed “Word Processing: Page Setup and File Edition.” The duration of the study was a 12-week period with 3 hours for each week. The participants performed portfolio creation, inspection and self- and peer-assessment via the Web-based portfolio assessment system developed for this study. The students didn’t have to create their portfolios by using presentation (e.g. Powerpoint) or Webpage (e.g. Frontpage) production software, rather by selecting the entries and filling out the forms in the Web-based portfolio assessment system to produce their e-portfolios.

2.2 Procedure of Experiment

The experiment included course unit 1 and course unit 2 with an artifact for each unit. Each course unit took 6 weeks. To elevate the reliability and validity, the experience of course unit 1 is particularly crucial for successful unit 2 in a way that increases grading skills and improves raters’ familiarity and shared perception of the assessment rubric. This study employed a full-blown and well-tested portfolio assessment activities designed by Chang and Tseng [6], and the course activities are illustrated as below:

1) The teachers, i.e. 1 instructor and 3 online assistants, demonstrated system operation and offered guide for students by delineating the assessment rubrics, scoring methods and scoring criteria. The well-trained online assistants were the

research members at this study who are fairly familiar with the assessment rubric and have a shared perception of the scoring standards.

2) Outside the classes, the students were involved in a number of course activities and online discussions. The activities for each course unit were: goal setting, reflection writing, artifact submission (including preliminary, revised, final versions of artifact), etc. The teachers viewed students' learning processes through the assessment system. The end-of-course examination score was determined by the average of a student's mid-term and final exam grades. The teacher, having over 10-year teaching experience, prepared the test forms for both exams.

3) Until each course unit ended, the teachers began to assess students' portfolios; in the meantime, students performed self- and peer assessment. Prior to the evaluation, the 4 teachers had reached consensus based on assessment rubric and scoring criteria. The whole class was divided into 12 groups in which each group member anonymously assessed 6 portfolios from the other group. Thus, each teacher was responsible for 72 portfolios, while each student had to assess a total of 7 portfolios (1 personal and 6 peer portfolios).

In the section of "Portfolio Assessment", teachers had access to a set of options by clicking on student names. These options were: Profile, Learning Goal, Reflection, Artifact, Other Entries, Scoring, Teacher Feedback, Peer Feedback, Participation Record, etc. By selecting "Scoring" option, teachers were allowed to report scores and write feedback. Students' portfolio contents and participation records could serve as a reference for teachers in their evaluation.

2.3 Development of Assessment Rubric

Our rubric was developed after reviewing the literature and discussing with the instructor in order to construct the face and content validity. The meetings with another three experts also helped to finalize the rubric and to establish expert validity. The rubric comprised 6 aspects with a total of 27 items, which were *portfolio creation, learning goal, artifact, reflection, attitude and other entries.* Scores were given ranging from 1 to 5, with a 0.5-increment in order to precisely distinguish the assessment results. In each rubric, various levels of performance were defined for precise scoring.

In this study, we adopted a scoring method suggested by Reckase [7], in which assessment results were converted to a hundred-mark system. To calculate a portfolio, original score was divided by the total score (160) with the quotient multiplying 100. Students' portfolio results can be a reference for teachers in determining semester grades.

2.4 Item Analysis of Assessment Rubric

The assessment rubric was firstly measured using item analysis. The t-value between high-scoring (27%) and low-scoring groups (27%) for each rubric aspect achieved significance level, which implied each rubric aspect had good discrimination capability and should be reserved. The Pearson's correlation between each rubric mean and overall mean was significant, showing that the consistency among rubric aspects was acceptably high and should be reserved.

2.5 Validity of Assessment Rubric

The Kaiser-Meyer-Olkin (KMO) values for each rubric aspect were greater than 0.5, implying that factor analysis could be applied. An approach of factor analysis─Principal Factor Analysis (PFA) ─ may be further used to construct validity. Considering that all factors (or aspects) had certain degree of correlation with one another, the oblique rotation approach was used conducting the PFA. The Chi-square approximate value of the Bartlett's test reached significance level. This finding confirmed the existence of common factors between the rubric aspects, which showed the applicability of factor analysis. The results of factor analysis indicated factor loadings of one rubric in the Attitude aspect were lower than 0.3; the rubric was thus neglected. This was probably due to the reason that the attitude rubric asked about students' opinions on the computer course, instead of on the portfolio assessment as other aspects did. However, the results of the second factor analysis revealed that all rubric aspects yielded factor loadings greater than 0.3. Therefore, all aspects were kept.

Five aspects with eigenvalues higher than 1 were refined. The overall explained variance was up to 90%, indicating the overall scale had high validity. The explained variances of each aspect were greater than 60%; this suggested each aspect was valid and effective in investigating the quality of portfolio contents. Among all aspects, Attitude yielded the highest explained variance followed by Learning Goal, whereas Portfolio Creation held the lowest variance. This demonstrated that Attitude was most likely to examine the portfolio content quality, while Portfolio Creation was not.

2.6 Reliability of Assessment Rubric

The rubric had high reliability with an overall Cronbach's α greater than 0.7. All aspects were also higher than 0.7, revealing that the rubric had a high degree of internal consistency. The value for Artifact was the highest, while that of Reflection was the lowest.

3 RESULTS

3.1 Reliability of Portfolio Assessment

Research question 1: Are the portfolio assessment results consistent among distinct raters?

Table 1 summarized the Pearson's correlation of 72 portfolio assessment results determined by 4 teachers (i.e. 1 instructor and 3 assistants). The results were highly correlated as well as significant. The hypothesis in the study was accepted. In other words, assessment results across 4 different teachers yielded a high degree of consistency. The greatest correlation coefficient appeared in the aspect of Portfolio Creation with the lowest in Reflection, indicating teachers were unlikely to reach consistency in the latter aspect. Although subjective judgment tends to affect assessment reliability [8], the findings in this study showed a strong correlation between evaluations based on 4 teachers. A comparable study result was elicited by Gelinas [9], who found out there was a high overall consistency between portfolio raters. In the reliability research of Gadbury-Amyot et al. [10], the correlation coefficients between 2 raters were measured ranging from 0.28 to 0.6; Rees and Sheard38 calculated the values between 0.36 and 0.69 ─ both study results were significant but lower than those of this study.

Table 1. Pearson's correlation of assessment results among distinct raters

Portfolio assessment	Correlation coefficient	Significance
Portfolio Creation	1.00	0.000***
Learning Goal	0.98	0.001***
Artifact	0.84	0.019**
Reflection	0.75	0.034**
Attitude	0.94	0.006***
Other	0.93	0.000***
Overall	0.91	0.000***

***p<0.01, ***p<0.001*

Research question 2: Are the portfolio assessment results consistent within individual rater?

The utilization of homogeneity coefficients is to investigate the consistency of an individual rater on different items, which was employed at this study in order to measure the consistency of a rater on distinct portfolios. The test for homogeneity evaluates the equality of several populations of a single categorical data by asking whether two or more populations are equal or homogeneous in some characteristics [11]. Therefore, the test of homogeneity of percentages was used to determine whether all the four raters (i.e. 1 instructor and 3 online assistants) demonstrated consistency in assessing different portfolio scores represented in the mode of percentage. The homogeneity analysis regarding the four raters is shown in Table 2. The homogeneity coefficients reached significance level, meaning each of the 4 teachers demonstrated consistency in terms of assessment results (i.e. inner-rater reliability). Likewise, the overall homogeneity coefficients were significant, implying the teachers were highly consistent in assessment results. The hypothesis in the study was accepted.

Table 2. Homogeneity coefficient of assessment results within individual rater

Faculty	Individual coefficient	Individual Z-value	Overall coefficient	Overall Z-value
Instructor	0.97	1.91**	0.95	1.87**
Assistant A	0.93	1.83**		
Assistant B	0.96	1.89**		
Assistant C	0.94	1.85**		

*** p <0.01*

3.2 Validity of Portfolio Assessment

Research question 3: Is each rubric appropriate for examining learning achievements, or do portfolio scores match well with student achievement test scores?

The Table 3 showed the coefficients of Pearson’s correlation between portfolio scores and student achievement test scores. The result revealed that these two variables were not only correlated but also significant. The hypothesis in the study was accepted. This implied consistency indeed existed; thus, portfolio rubrics were appropriate for detecting students’ learning achievements. Gelinas [9] made a statement in accordance with this finding: there is a positive correlation between learners’ portfolio scores and their academic performance. Besides, “Learning Goal” and “Portfolio Creation” had higher possibility to reach consistency, for they were the two aspects with highest correlation coefficients.

Table 3. Pearson’s correlation between portfolio scores and achievement test scores

Aspect	Correlation coefficient	Significance
Portfolio Creation	0.78	0.000***
Learning Goal	0.79	0.000***
Artifact	0.67	0.000***
Reflection	0.63	0.000***
Attitude	0.59	0.000***
Other	0.57	0.000***
Overall	0.84	0.000***

****p<0.001*

4 DISCUSS AND CONCLUSION

Based the finding above, Web-based portfolio assessment may be considered as reliable and valid. This conclusion coincides with the belief of other researchers. Nevertheless, opposite results were discovered by Derham and Diperna [4], who demonstrated contradictory findings in their reliability study of digital portfolios, in which correlation coefficients between 2 raters were weak and insignificant. They also figured out low and insignificant correlation coefficients in the validity research on digital portfolios. There are a number of factors that may be responsible for divergent research findings, including students' backgrounds, sample sizes, the subject matter involved, digital portfolio environment, assessment rubrics, rating trainings, and scoring standards adopted. Oskay, Schallies and Morgil [12], after reviewing relevant investigations, asserted that portfolio assessment is not only valid but reliable. Although disagreement can be found in past studies, most researchers treated portfolio assessment as a tool with reliability and validity. Even though low level of reliability was discovered, Sulzen, Young and Hannifin [13] concluded increasing the number of raters was effective in reliability improvement. That is to say, it is not impossible having portfolio assessment reliable as well as valid.

All Cronbach's α values of each rubric aspect exceeded 0.7; this signified portfolio rubrics had sufficient reliability. Among which, the highest value was measured in "Artifact" and the lowest in "Reflection". In inter-rater reliability analysis, a high level of consistency appeared among 4 teachers in terms of assessment results. Likewise, in inner-rater reliability analysis, each rater was found highly consistent in scoring portfolios. Consequently, Web-based portfolio assessment may be a reliable approach that possesses both inter- and inner-rater reliability. In addition, many researchers indicated a number of factors that could ensure scoring reliability, including well-trained raters, concrete assessment rubrics, or raters' common perception about scoring criteria [4, 10, 12]. Sulzen, Young, and Hannifin [13] further pointed out that expanding the number of raters is also considered as instrumental in promoting reliability. This study attempted to provide assessors with opportunities for discussion and to help them become familiar with the rubric prior to the evaluation of portfolios. As a result, not only the inter-rater but also the inner-rater reliability was found to be significant. In response to reliability issues, we believe that raters are supposed to grade in an impartial way and have deep understandings and common perceptions about assessment rubrics.

In factor analysis, the overall explained variances maintained high, which meant rubrics of Web-based portfolio assessment should be useful in measuring the quality of portfolios. Also, rubrics were capable of mirroring certain degree of learning achievements in light of the strong consistency observed between students' portfolio scores and achievement test scores. Given these findings, Web-based portfolio assessment and its rubrics had good level of validity. Chang [3] noted assessors, during assessment process, must scrutinize various aspects of students' learning. Therefore, it is concluded that assessment rubrics are preferably multifaceted and full-blown so that learning achievements can be faithfully evaluated.

REFERENCES

[1] 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.

[2] Lu, C.-Y. (2001) Let students be the master of learning: Application and reflection of portfolio assessment. *Magazine of Tunhuang English teaching, 30,* 15-21.

[3] Chang, L.-L. (2002). Reliability and validity analysis of portfolio assessment: A case of composition portfolio in an elementary school. *Research in Education and Psychology, 25*, 1-34.

[4] 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.

[5] 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.

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

[7] Reckase, M. A. (2002). *Portfolio define*. Paper presented at the Workshop of Portpolio Assessment, Taipei, Taiwan.

[8] Yu, M.-N. (2003). *Educational test and assessment*. Taipei, Taiwan: Psychology Publisher.

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

[10] 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.

[11] Chiu, H.-C. (2009). Quantitative research and statistical anslysis. Taipei: Wu-Nan.

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

[13] 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.