Assessing E-Portfolios of Practical Production: An Investigation of Marking Methods in Visual Arts and Design

Hendrati Nastiti[1], Paul Newhouse[2], Jeremy Pagram[3]
[1,2,3]Edith Cowan University (Perth, Australia)
[1] hnastiti@our.ecu.edu.au

ABSTRACT

This paper reports on a study that was a part of a three-year project which was a collaboration between Edith Cowan University and the School Curriculum and Standard Authority (SCaSA) of Western Australia under the funding by Australian Research Centre (ARC) Linkage Program, Edith Cowan University and the SCaSA of Western Australia. The focus of this paper is to discuss the validity of comparative pairs judgments as an alternative scoring method for summative assessment of practical production performances. The comparative pairs judgments method has been considered to have the potential for a more holistic and objective scoring of students' achievement than the traditional analytical method of assessment. The sample for the study consisted of 157 secondary school students in either the Visual Arts or Design courses, and the total of 20 experienced assessors. Through this investigation the study sought to engage in a discussion of assessment methods that are authentic, formative, and multimodal; as have been made possible by the development of Information and Communications Technology (ICT).

KEYWORDS

Comparative Pairs, digital assessment, scoring methods, validity, reliability

1 INTRODUCTION

Issues and problems in assessment, particularly in courses with major practical components, have made it essential for educators to consider alternative assessment methods that could produce results that are reliable and accountable. This study investigated the validity of the scores generated from the comparative pairs judgments method, based on Thurstone's "Law of Comparative Judgment" [1], in assessing student practical work in the Visual Arts and Design courses.

The development of Information and Communications Technology (ICT) has made the assessment of student work using the comparative pairs judgments possible. With the current development of ICT, computer software could be developed to generate the pairings in a matter of seconds, and the scoring result could be calculated, viewed and analysed instantly. Taking the advantage of the development in ICT, educational researchers have experimented on, and made advances in, the use of the comparative pairs judgments for assessing student achievement on different courses such as English, Engineering, Italian and Physical Education [2, 3].

This paper discusses the validity of the comparative pairs judgments method in the first phase of a three-year project conducted by the Centre for Schooling and Learning Technologies in Edith Cowan University in Perth. The project investigated the challenges of the comparative pairs judgments method in two secondary school courses: Design and Visual Arts.

2 COMPARATIVE PAIRS JUDGMENTS VERSUS ANALYTICAL MARKING

The comparative pairs judgments method was based on Thurstone's Law of Comparative Judgment [1] and was further developed by Alistair Pollitt and David Andrich [2, 4, 5]

based on the Rasch logistic model. In this method the assessors judge the better works between pairs of works based on a holistic criterion instead of assigning scores in an outcome-specific assessment rubric in analytical marking. From these judgments, an interval scale is generated to represent the quality of the work compared to the other works [5].

In his paper, Thurstone [1] argued that a holistic judgment of pair comparisons is a better way to measure latent traits. Pollitt [5] further elaborated on the reliability and validity of the comparative pairs judgments method, especially compared to the commonly and widely used analytical marking.

In analytical marking, the scoring is guided by an assessment rubric. The purpose of this rubric is to control subjectivity [6, 7], and consequently to increase the reliability of an assessment, especially in practical assessment. However, in this type of assessment, subjectivity often remains a problem. Assessors' personal standard and interpretation could still influence their judgment in assigning scores for the outcomes. When the scores are added up, this factor would also add up, resulting in a total score that is likely to include each assessor's subjectivity.

In the assessment rubric used in Analytical marking, each criterion is broken down into levels of achievement with a defined weighting. Several issues could be identified from this scoring process. Firstly, with a scaled score, assessors tend to mark a work between the minimum and the maximum score. For example, if the score range is from 1 to 6, assessors tend to give a score of 2, 3, 4 or 5. This means that the minimum and maximum scores are often not utilised. When the scores are added up, the underutilisation of the minimum and maximum scores becomes more obvious.

Secondly, with a range of scores, the assessors' personal standard and how they identify different skill levels with the scores within the range would create a difference among assessors [5, 8]. This would also become more obvious when in the total score.

These two problems do not exist in comparative pairs judgments because there are only judgments of the better work out of a pair instead of assigning scores. Pollitt [5] recognised two problems that would also exist in comparative pairs judgments, the first of which is the variation in ranking among assessors. The second problem is, even though the comparative pairs judgments uses a holistic criterion, it still consists of different skill levels that make judging the better quality of work a complex process.

From the point of view of the sources of unreliability of the scoring result, comparative pairs judgments appears to have less problems. This is also supported by several research studies [3, 9, 10, 11] that have been done on this alternative method with a high reliability. Because of the positive overview of the comparative pairs judgments, this method was suggested to be used for high-stakes practical assessment [5, 10].

The Project

In 2011 the Centre for Schooling and Learning Technologies in Edith Cowan University in Perth started a three-year project titled Authentic Digital Representation of Creative Works in Education. This project was a collaboration between the School Curriculum and Standard Authority (SCaSA) (formerly the Curriculum Council) of Western Australia and funded by Australian Research Centre (ARC) Linkage Program, Edith Cowan University and the SCaSA of Western Australia.

The aim of the project was to "investigate the representation of student practical work in digital forms for the purpose of summative assessment and online marking using the comparative pairs method" [3]. This project was conducted in three phases, with each phase was run in one year. The first phase was the pilot project in which the researchers digitised the student work. In the second phase the students digitised their own work and was focused on the feasibility of the complete

digital assessment process. The third phase was focussed on the scoring methods.
This paper reports on a part of the first phase of the main project that focuses on the reliability and validity of the scores generated from the comparative pairs judgments method.

3 THE ASSESSMENT SYSTEMS

The study focussed on the use of an assessment management system that was specialised in the Comparative Pairs marking. The system was called the Adaptive Comparative Judgment (ACJ) system. It was developed within the Technology Education Research Unit (TERU) project at the Goldsmith College University of London. This web-based assessment system was designed to save the digital copy of student work in a database, create the pairings of student work, display these pairings for marking, and provide statistical analyses of the marking result and the judgment process. The system has been trialled in several institutions in several countries such as UK, Singapore, Sweden and Spain.
The ACJ system was designed to dynamically pair the portfolios of student work output based on previous judgments. The judgments are grouped in judgment rounds. In the first round the pairing is done randomly, resulting in 50% 'winners' and 50% 'losers'. In the second round the ACJ system paired portfolios within the two groups, resulting in three groups which consisted of works that have never won, won once, and won twice. Pairings for the third round were created among works within the three groups, and so the system continued, until there was enough information for the Rasch parameters to be established.
The ACJ system then creates pairings that "will provide the most information for increasing the reliability of the rank order" [11] by putting up pairs of work that were of more and more similar nature. Because of this adaptive function this judgment is also called "Adaptive Comparative Judgment".
Starting from the seventh round, a different pairing method is used. In this pairing method, 'chained' pairing is used. One student work from the first pair within a group was kept for the next pairing to be compared with another work. This was considered to make judging easier for judges and increase the efficiency of the judging process [2]. After each round of judgments, the system analyses the data resulting from the judgments, including the location of the portfolio relative to the other portfolios and assessor misfit statistics. From this analysis the inconsistencies in the judgment may be detected early. If after a round the reliability coefficient was still considered not sufficiently high, another round was created. Once a reasonably high reliability had been achieved, the marking process was considered finished and data were processed to be analysed. The reliability coefficient calculated represented both the internal reliability and the inter-rater reliability.

4 METHODOLOGY

The research methodology used in this paper is the mixed research methodology. Data analysed were quantitative data obtained from the scoring results and qualitative data from interview with the assessors. The first phase of the project that is reported in this paper was conducted concurrent to the Western Australian Certificate of Education (WACE) examination with year 12 students in Western Australia.
For WACE examination, year 12 students who were studying Stage 3 Design and Visual Arts courses submitted their practical assessment for marking. In the Design course, the submitted work was in the form of a 15-page single-sided portfolio consisted of examples of their Design projects. In the Visual Arts course, the students submitted a resolved artwork accompanied by an artist statement. The portfolios were scanned by the research team and saved as a pdf file. The artworks were photographed and video-recorded. The digital representation of the

student work was then uploaded into the database on the assessment systems for scoring.

- In Design there were 82 students from six schools in Western Australia were involved in the main project. Ten experienced assessors were involved in the comparative pairs judgments and two were involved in the Analytical marking.
- In Visual Arts there were 75 students from ten schools in Western Australia involved in the main project. Fifteen experienced assessors were involved in the comparative pairs judgments and three were involved in the Analytical marking.

For this study the assessment criterion for comparative pairs judgments was a holistic criterion derived from the assessment rubric used in the other two marking processes. This holistic criterion was a summarised version of the criteria in the rubric, but without the weighting that the rubric had. The criterion was discussed and agreed upon by the assessors in each course.

As a comparison, marking results from two other scoring processes were used. One set of result was from the Analytical marking conducted within the project and the other set was from the WACE practical marking conducted by assessors in SCaSA. These two processes were quite similar, with the difference was only on the type of work being marked. In the Analytical marking process the assessors marked the digitised work, which was the same digitised work used in the comparative pairs judgments. In the WACE marking the assessors marked the original work submitted for the examination. The Analytical marking process was conducted on a Filemaker Pro database that was developed especially for the main project. Table 1 below displays the nature of each scoring process.

Table 1. Comparison of scoring methods

	CP	Analytical	WACE
Assessment Criteria	Holistic	Rubric	Rubric
Type of work	Digital	Digital	Design: Portfolio VA: Original artwork
Number of assessors	10	2-3	> 2

Note: CP=Comparative Pairs judgment, WACE=the practical component of the Western Australian Certificate of Education examination

5 FINDINGS

In this study, the validity of the comparative pairs judgments method of scoring is discussed from three points of view. The first point of view is the reliability of the result of judgment as one measure on validity [10]. The second is the comparability of the result of judgment with the scores obtained from the other two marking processes [8, 12]. The third is a discussion on the issues that might reduce the validity of the judgment result [13]. As the result of the comparative pairs judgments was in a rank order while results from the other two scoring methods were in interval scale, the comparative pairs judgments score in this analysis was the result from rescaling using result from the analytical marking as a standard.

On the issue of reliability, both the Analytical marking process and the comparative pairs judgments method had high reliability coefficients, as is shown in Table 2 below.

Table 2. Internal reliability for each set of scores

Judgment method		Internal reliability	
		Design	Visual Arts
Analytical marking	A1	.953	.934
	A2	.950	.915
	A3	n/a	.934
	Ave	.962	.944
CP		.941	.959
WACE		n/a	n/a

Note: A=Assessor, Ave=Average, CP=Comparative Pairs judgment, WACE=the practical component of the Western Australian Certificate of Education examination

The high internal reliability for both analytical marking assessors represented the internal reliability of the criteria. The coefficient indicated that there was an overall agreement among the criteria in the rubric. The inter-rater reliability was represented in the correlation between assessors.

Even though the internal reliability for the Analytical marking was high, the correlations between the scores generated by analytical marking assessors were only moderate, with correlation coefficients ranging from .51 to .56 ($p<.01$) in both courses, as shown in Tables 3 and 4 below. These coefficients indicated that there was only moderate agreement between assessors in the Analytical marking even though the assessors were using the same assessment rubric. This highlighted the concern over the reliability of the marking of subjective courses such as Design and Visual Arts. This issue was consistent in the two courses; therefore there was nothing that suggested the difference in the type of the assessment task made a difference in the agreement or disagreement between assessors.

Table 3. Correlations between scores from the three methods of scoring in Design

(N=82)	A1	A2	Ave	CP	WACE
A1	1	.53**	.89**	.61**	.55**
A2		1	.86**	.48**	.36**
Ave			1	.63**	.52**
CP				1	.67**
WACE					1

Note: **. Correlation is significant at the .01 level (2-tailed). A=Assessor, Ave=Average, CP=Comparative Pairs judgment, WACE=the practical component of the Western Australian Certificate of Education examination

Table 4. Correlations between scores from the three methods of scoring in Visual Arts

(N=75)	A1	A2	A3	Ave	CP	WACE
A1	1	.54**	.51**	.84**	.68**	.70**
A2		1	.56**	.82**	.72**	.75**
A3			1	.83**	.58**	.71**
Ave				1	.79**	.86**
CP					1	.74**
WACE						1

Note: **. Correlation is significant at the .01 level (2-tailed). A=Assessor, Ave=Average, CP=Comparative Pairs judgment, WACE=the practical component of the Western Australian Certificate of Education examination

While for the Visual Arts course the correlations between WACE practical examination scores and the other scoring processes were all similarly strong, it was not the case for the Design course. In the Design course, the Comparative Pairs scores were moderately correlated to the WACE scores but the Analytical marking score from the Assessor 1 was only moderately correlated to the WACE scores and the scores from Assessor 2 were only weakly correlated to the WACE scores despite the two scoring methods utilised the same rubric. This further suggested that there might be a problem with the rubric and how the assessors interpreted the rubric.

Several factors could have influenced this weak-to-moderate correlation between scores from the Analytical assessors and the WACE

practical examination. Firstly, the WACE in 2011 was only the second year that Design was among the subjects assessed, and because of that the assessment rubric might not have been examined too well. Secondly, the varied assessors' design teaching specialisations could be a problem, since they came from specific different design background for example photography or technical graphic while the student work could be in any Design strand. That in the Design WACE examination score reconciliations were needed because for a number of students the scores were too different (A. Price, personal communication, June 14, 2012) might be as a result from these two factors.

Parallel to this, one of the findings for Design for the project was on the under-utilisation of some of the score range of the criteria. In the project it was found that for some criteria, the lowest and highest ends of the score range were not used by the assessors, especially by Assessor 2 [3]. The other possible factor was the difficulty in marking the Design portfolios. In Design, the task was a 15-page portfolio consisted of evidence of design processes while the rubric consisted of six criteria with a score from six to ten for each criterion. The size of the work, combined with a wide range of score for each criterion opened the possibility of errors in scoring.

These problems highlighted the importance of the quality and appropriateness of the assessment rubric in Analytical marking. The assessment rubric needs to be rigorously developed and tested in order for it to provide guidance for reliable marking.

The reliability coefficient of the comparative pairs judgments, on the other hand, represented both the internal reliability, or internal consistency in judgment, and the inter-rater reliability [2]. As was shown in Table 2, this reliability coefficient was high. The judgments in the ACJ system were stopped when the reliability coefficient of the result was considered sufficiently high, which occurred on the thirteenth round in both Design and Visual Arts courses, with coefficients of .941 and .959 consecutively. In the thirteen rounds there were 543 judgments made in Design and 497 judgments made in Visual Arts.

Beside the reliability coefficient, the system also reported the misfit statistics, which could reveal any inconsistency in judgment and among assessors. At the end of the thirteenth round in the comparative pairs judgments method there was only one assessor who showed inconsistency with other assessors in Design course and there were none in Visual Arts.

From the interview with the assessors, there were several factors that the assessors were concerned about. In Design, the quality of the digital representation was considered to be adequate, except when the original was drawn with pencil and could not be scanned well. Both assessment systems were reported to be easy and convenient to use. There was a mixed response on the marking methods and the criteria used in the marking methods. Assessors who favoured the Comparison Pairs judgment considered the method to be more accurate, straightforward and objective, and the holistic criterion made judging easier. Assessors who favoured the Analytical marking with an assessment rubric considered the method to allow for a more careful, specific, accurate and accountable.

In Visual Arts, on the contrary, most assessors reported that the digital representations did not adequately represent the original work, especially in certain types of material and medium. However, in general, the assessors did not report major problems with the assessment systems aside from problems that were related to the quality and limitations of the digital representations. On the judging experience, many assessors reported that the problems they encountered were mainly caused by the limitations of the digital representation. One assessor preferred analytical marking because the marking rubric provided a good guideline while another preferred the comparative pairs judgments method because it involved more

assessors and considered this method was easier especially when the quality of the digital representation was low.

CONCLUSIONS

This study investigated the validity of the comparative pairs judgments method as an alternative scoring method for summative assessment of practical production. As a comparison marking results from digital Analytical marking and the WACE practical examination marking were used. The investigation was conducted in two secondary school courses, Design and Visual Arts.

In general, both the internal reliability and inter-rater reliability in the comparative pairs judgments scores were high in both courses with none to very few inconsistencies in judgment analysed. When compared to the WACE practical examination result, results from the comparative pairs judgments were also strongly correlated. Conversely, findings from the analytical marking suggested that while assessment rubric might reduce the subjectivity in the scoring of practical assessment task, it has its limitations. Since Analytical marking depends on both the quality of the assessment rubric and the assessors' personal standard, an assessment rubric needs to be rigorously examined for it to provide assessment result that is reliable, valid, equitable and accountable. This problem does not exist in comparative pairs judgments method.

There did not appear to be a big difference between how the type of task in the two courses affected the validity of the scores from the comparative pairs judgments method. In both courses the high reliability and the correlation with the WACE examination result was strong. As the interview with the assessors indicated, however, the Visual Arts assessors were more concerned over the limitations of the digital representations than the Design assessors. While in Design course the quality of the digital portfolio were close to the quality of the original work, capturing the quality of the artwork in Visual Arts course was more complicated. This could consequently affect the validity of the scoring result in Visual Arts course.

As is in analytical marking, in comparative pairs judgments the assessors' knowledge, skill and experience are important. In comparative pairs judgments the assessors used a holistic criterion to judge the 'winner' within a pair of portfolios, consequently this opens to the possibility that the judgment is made based more on the appearance of the portfolios instead of other qualities.

Interview with assessors from both courses indicated that the quality of the digital representations needed to be improved. Managing a large number of high quality digital files is still a problem despite the current advancement in ICT. If assessment is to turn digital, educators need to optimise the whole assessment process without compromising the quality of the digital representations of student original work to maintain the validity of the assessment result.

While comparative pairs judgments showed good reliability and validity for practical summative assessment as was reported in this study, it might not be sufficient for other types of assessment. For example, the result of this judgment is in a rank order; therefore it may not be an appropriate method for mapping individual student's performance such as in formative or evaluative assessment.

REFERENCES

[1] Thurstone, L. L. (1927). A law of comparative judgment. Psychological Review, 34, 14.

[2] Kimbell, R. (2008). E-assessment in project e-scape. Design and Technology Education: An International Journal, 12(2).

[3] Newhouse, C. P., Pagram, J., Paris, L., Hackling, M., & Ure, C. (2012). Authentic digital representation of creative works in education: Addressing the challenges of digitisation and assessment. Perth, Western Australia: Centre for Schooling and Learning Technologies (CSaLT) Edith Cowan University.

[4] Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. Applied Psychological Measurement, 2(3), 451-462.

[5] Pollitt, Alastair. (2012). The method of Adaptive Comparative Judgment. Assessment in Education: Principles, Policy & Practice, 19(3), 281-300.

[6] Mertler, Craig A. (2001). Designing scoring rubrics for your classroom. Practical Assessment, Research & Evaluation, 7(25), 1-10.

[7] Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. Practical assessment, research & evaluation, 7(10), 71-81.

[8] Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. Educational Research Review, 2(2), 130-144.

[9] Heldsinger, S., & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. The Australian Educational Researcher, 37(2), 1-19.

[10] Kimbell, R., Wheeler, T., Stables, K., Sheppard, T., Martin, D. Davies, Pollitt, Alistair, & Whitehouse, G. (2009). E-scape portfolio assessment: phase 3 report. London, UK: Technology Education Research Unit, Goldsmiths, University of London.

[11] Whitehouse, C., & Pollitt, A. (2012). Using adaptive comparative judgment to obtain a highly reliable rank order in summative assessment. Retrieved June 11, 2013 from https://cerp.aqa.org.uk/research-library/using-adaptive-comparative-judgment-obtain-highly-reliable-rank-order-summative-assessment

[12] Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research & Evaluation, 9(4), 66-78.

[13] Shaw, S, Crisp, V, & Johnson, N. (2012). A framework for evidencing assessment validity in large-scale, high-stakes international examinations. Assessment in Education: Principles, Policy & Practice, 19(2), 159-176.