2018

Comparative pairs judgements for high-stakes practical assessments

Hendrati Nastiti
Edith Cowan University

Comparative Pairs Judgements

for

High-Stakes Practical Assessments

This thesis is presented for the degree of

Doctor of Philosophy

Hendrati Nastiti

Edith Cowan University

School of Education

2018

USE OF THESIS

DECLARATION

I certify that this thesis does not, to the best of my knowledge and belief:

i. incorporate without acknowledgment any material previously submitted for a degree or diploma in any institution of higher education;

ii. contain any material previously published or written by another person except where due reference is made in the text of this thesis; or

iii. contain any defamatory material.

I also grant permission for the Library at Edith Cowan University to make duplicate copies of my thesis as required.

Signature: [redacted]

Date: April 2018

ACKNOWLEDGEMENT

I would like to express my gratitude and deep respect to my supervisors, the late Associate Professor Paul Newhouse and Dr. Jeremy Pagram who guided me and believed in me, especially at times when I did not believe in myself. Paul and Jeremy's knowledge, perception, hard work, persistence, and passion in education have made this thesis possible. They gave me the opportunity to develop myself and provided me with a learning experience far beyond what I could imagine.

My thanks to ECU staff who have helped and encouraged me along the way: Dr. Alistair Campbell, Dr. Martin Cooper, Dr. Pina Tarricone, Zina Cordery, Bev Lurie, Dr. Yvonne Haig, Dr. Julia Morris, Dr. Jan Gray, Dr. Bill Allen, Dr. Geoff Lowe, Sarah Kearn, and the IT staff.

My thanks also goes to my friends whose understanding and support have lifted up my spirit during the difficult periods: Dr. Beatriz Cuesta-Briand, Dr. Dawon Seo, Lim Mei Lian, Dr. Marina Confait, Michelle Tan, Sarah Booth, Dr. Mayada Mhanna, Talisha Goh, Jonathan Messer, Jennifer Jin, Lan Thi Thu Nguyen, Hiep Vu Thi Bich, Ngan Phan Thu Vo, the McKimmies, and my badminton team.

Without the encouragement, help, and understanding of my family, relatives, friends, and especially my supervisors, I would not be able to finish this thesis.

ABSTRACT

Assessment of practical tasks, as opposed to that of theoretical tasks, has been considered to be problematic, mainly because it is usually resource intensive and the scoring is subjective. Most practical tasks need to be assessed on site or involve products that need to be collected, stored, or transported. Moreover, because practical tasks are generally open-ended, and therefore subjective, there is concern over the reliability of the scores. In high-stakes assessment, these problems are even more challenging. There is a need for an assessment method that could overcome these problems. In this study, such a method that will be referred to as the Comparative Pairs judgements was investigated. This scoring method was applied to samples from the practical examination in two secondary courses in Western Australia: Design and Visual Arts.

This study was conducted within the first phase of an Australian Research Council (ARC) Linkage Project titled the *Authentic Digital Representation of Creative Works in Education*. This main project was a collaboration between the Centre for Schooling and Learning Technologies (CSaLT) in Edith Cowan University and the Curriculum Council of Western Australia. The purpose of the present study was to investigate the suitability of the Comparative Pairs judgements as an alternative assessment method for assessing high-stakes practical production tasks. The overarching research question was *how representative are the Comparative Pairs judgement scores of the quality of the student practical production work in Visual Arts and Design courses?* In the present study, student work that was submitted for the practical examination was digitised for online scoring processes. The digital representation of student work enabled online access for judging, regardless of the location of the assessors. Both a Comparative Pairs judgements method and an Analytical marking method were used to score these digital representations.

An interpretive research paradigm was employed, by utilising an explanatory sequential mixed method design. Data collected for the present study were part of the data collected in the main project. While data for the main project was quite extensive, only scoring data and the assessor interviews and online notes were considered relevant to this study, and therefore only these data were analysed and discussed in this thesis. A total of 157 students studying Design and Visual Arts participated in the first phase of the main project and the present study. A total of 25 assessors participated in the Comparative Pairs judgements and the Analytical marking processes.

Scoring data analysed in this study were obtained from three scoring processes: the official practical examination scores, the online Analytical marking, and the Comparative Pairs judgements. Data analysis included descriptive statistics, correlation analysis, Rasch dichotomous modelling, fit statistics, and reliability analysis. A further discrepancy analysis was conducted on student works that showed scoring inconsistency, either between methods of scoring or between assessors. Data from the assessor interviews and judgement notes from the scoring processes were triangulated with the scoring data to examine the validity of the Comparative Pairs judgements method as an alternative scoring method. Data from the scoring of the digital representations of the student work in Design and Visual Arts were analysed separately to examine the suitability of the Comparative Pairs judgements in each course, and consequently compared to examine the influence of the different assessment tasks in the two subjects on the scoring result.

Findings for both the Design and Visual Arts courses suggested that the scoring resulting from the Comparative Pairs judgements was reliable. This was mainly due to the numerous judgements and the pairing algorithm, therefore the inconsistencies in judgements were cancelled out, creating scoring results that could be more reliable than the more commonly used Analytical marking. The validity analysis that was conducted used both the evidence for, and threats against validity, suggested that this assessment method could be a valid method for high-stakes practical assessment in these two courses.

The present study found that the reliability of the scores and the validity of the Comparative Pairs judgements as an assessment method make this method suitable for assessing high-stakes practical production. Findings from the present study suggested that this method is applied and further investigated in different educational settings for

different practical assessment tasks. This method of judgements should be considered to be potentially valuable for formative assessment and summative assessment alike, as well as teacher professional learning, and moderation practices.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

TERMS AND ABBREVIATIONS

ACARA	Australian Curriculum, Assessment and Reporting Authority
ACJ	Adaptive Comparative Judgement
ARC	Australian Research Council
CC or	Curriculum Council
SCSA	School Curriculum and Standards Authority, the institution was previously named Curriculum Council of Western Australia. The change to the current name happened during the writing of this thesis, therefore the name SCSA is only used for discussions after the data collection.
CTT or	Classical Test Theory or
TTT	Traditional Test Theory
CP	Comparative Pairs
CRT	Criterion-Referenced Test
CSaLT	Centre for Schooling and Learning Technologies
ECU	Edith Cowan University
ICT	Information Communication Technology
IRT	Item Response Theory
NRT	Norm-Referenced Test
PDF	Portable Document Format
PPT	PowerPoint
TERU	Technology Education Research Unit (Goldsmiths College: University of London)
TPACK	Technological Pedagogical Content Knowledge
TTT	Traditional Test Theory

WACE	Western Australian Certificate of Education
wms	Weighted mean square

CHAPTER 1
INTRODUCTION

The use of the Information and Communications Technology (ICT) in education as a medium for teaching and learning as well as for collaboration among teachers, students, parents, school administrators, and education policy makers has become increasingly common and expected in schools around the world (Blanchard & Moore, 2010). The use of ICT to assess student achievement, however, is less common, especially in high-stakes assessment (Miller, 2011; Timmis, Broadfoot, Sutherland, & Oldfield, 2016). Aside from using computers to replace pen and paper, most current high-stakes assessment practices still use traditional assessment methods, including for practical tasks. Currently, most academic high-stakes assessments such as the SAT in the United States are still conducted on paper. Practical tasks in high-stakes assessment are either marked on site, for example the speaking component in the International English Language Testing System (IELTS), or sent to a marking venue, for example the Visual Arts practical component of the Western Australian Certificate of Education (WACE) examination in Western Australia. Some would suggest that there is a technology gap between teaching and learning practices and the associated assessment practices (Stables, 2017b; Timmis et al., 2016).

As educators have increasingly espoused constructivist views of learning, the interest in more authentic forms of assessment also grows (Wiggins, 2011). In constructivist learning, students construct their knowledge based on their existing knowledge and experience, as well as from their social and physical environment. They think, research, collaborate, discuss, design, create, and evaluate in tasks that are authentic and relevant to real situations (Binkley et al., 2012; Fullan & Langworthy, 2013). Examples of such tasks are portfolios, performances, research projects, and many others; often as a combination of different skills, topics, and subject areas; and researched, created, recorded, or manipulated using ICT. This approach to learning calls for more authentic forms of assessment that more naturally flow from these learning tasks (Binkley et al., 2012; Masters, 2013).

Because of its versatility, ICT can well facilitate more authentic forms of assessment (Masters, 2013). The use of ICT removes various limitations that previously constricted assessment processes, for example time and media limitations (JISC, 2010). It also creates possibilities previously unavailable such as learning from computer simulations or creating three-dimensional design prototypes. Students can now create, record, edit, and submit their work digitally through video recording, audio files, pictures, and digital portfolios (Timmis et al., 2016). Not only are these digital works flexible, they also could better represent student achievement than the traditional pen-and-paper test. In particular, digital technologies offer affordances that can support more authentic assessment of practical performance whether that be to produce an artefact such as artwork or perform human movement such as dance.

For school and assessment authorities, digital representation of performances could be recorded, judged, moderated, and reported online with ease (JISC, 2010). Digital representations can be cost-effective because they do not need physical storage or to be physically transported (Masters, 2013). As an example, students' artworks that could be in different shapes, materials and dimension could be photographed or video recorded and made available for online scoring instead of having to be sent to an examination venue. Recorded visual or aural performance such as music instrument skills and dance movement could be viewed at different times by different assessors and for different assessment purposes such as for online moderation (Adie, 2013). Assessors in different places could assess digital representations of student work across the country and around the world.

Different methods of assessment such as online marking and online moderation are now more easily facilitated (Adie, 2013; Jordan, 2013). Different measurement methods could be more easily used to judge student achievement. For example, the Comparative Pairs judgements method was not feasible to use for mainstream assessment until online tools were available (Whitehouse & Pollitt, 2012). Comparative Pairs judgements is a measurement method that was first introduced in the late 1920s, at the dawn of psychological and educational measurement practice (Thurstone, 1927). In the literature it may also be referred to as *Pairwise Comparison* (Heldsinger & Humphry, 2010) or *Comparative Judgements* (Pollitt, 2012b). Only recently has this method been extensively researched, largely because the current digital technologies could now support it. The use of the Comparative Pairs judgements method has since been investigated in countries such as the United Kingdom, Australia, USA, and Singapore. These studies indicated this measurement method to be able to provide an assessment result that is reliable and valid (Kimbell et al., 2009; Pollitt, 2012a, 2012b), especially for assessment tasks that are subjective in nature such as speaking and essay writing.

If the Comparative Pairs judgements method is to be used for assessing high-stakes practical assessment, it needs to be rigorously examined. Important factors such as the feasibility, acceptability, and appropriateness for different types of assessment tasks need to be thoroughly investigated. The feasibility factor of this judgements method includes the feasibility of the creation of the digital representation of student work, and the feasibility of the online scoring system, which includes the quality of the digital representation and the accessibility of the scoring system. Teachers, students, and the assessment authorities need to have evidence-based confidence in the quality of the judgement results for this method of measurement to be accepted. This judgements method also needs to be appropriate for the assessment task, which means that the reliability of the scores and the validity of the assessment need to sufficiently high (Moss, Girard, & Haniford, 2006). The present study investigated these factors in two Western Australian senior secondary school subjects with a major practical component: Design and Visual Arts. Throughout this thesis, *Design* with a capital D refers to the Design course and *design* in lowercase refers to the general terminology.

Background

The move from the industrial age to the information age has transformed the way people work and live (Griffin, Care, & McGaw, 2012). The rapid advance of technology has provided more powerful and affordable electronic devices such as mobile phones and tablets, and access to information and resources on the Internet. These digital technologies have opened up new possibilities (e.g., internet-based trading and online learning) and erased boundaries (e.g., collaborations among people from different places, generations, and occupations) (Binkley et al., 2012). ICT has traversed borders and connected homes, schools, industries, remote areas, and countries, opening up new ways to think, learn and create. Crowd sourcing and eCommerce are among the many web-based global activities made possible by ICT (Brabham, 2013; Martindale & Dowdy, 2016).

The connectedness of almost all aspects of life is often succinctly coined as the *Internet of Things*, which The International Telecommunication Union (ITU, 2012), a United Nation agency, defined as “A global infrastructure for the information society, enabling advanced services by interconnecting (physical and virtual) things based on existing and evolving interoperable information and communication technologies” (p. 1). This term describes the immensity of the use of various digital technologies in society. These uses have now included the systems used in factories to manufacture and track their products and assets, by farmers to oversee their cattle or produce, and in restaurants to manage orders and stocks. Collaboration between government agencies, organisations, industries, and institutions from different places and in various projects is also more feasible now.

Similarly, in education the availability of digital technologies has created new possibilities and opened up new opportunities. Digital technologies can be used to facilitate more flexible, contextual, and constructivist learning and teaching activities that were not available only a decade ago (Allison & Kendrick, 2015; Lim, Zhao, Tondeur, Chai, & Chin-Chung, 2013). In science, for example, students could now use technologies such as data loggers to measure variables such as speed or temperature in science projects and manipulate the data to understand concepts, create presentations, or draw conclusions more clearly than if they were conducted manually. In design and technology students could easily design furniture, houses, or machineries using software such as CorelDraw, SketchUp, and Autodesk. Students from different schools in different countries could communicate and collaborate in projects such as iEARN (International and Education

Resource Network, n.d.), BRIDGE (Asia Education Foundation, n.d.), and the BigDayta (BigDayta, n.d.).

Not only has ICT facilitated changes to teaching and learning in education, it has also opened up new demands on education systems. Job markets are changing; ICT skills have become a general requirement instead of a specialty. Creativity, critical thinking and problem solving skills are becoming more important (Binkley et al., 2012; Casimaty & Henderson, 2016). Employers look for employees who can collaborate and learn new concepts; who are creative and innovative (Griffin et al., 2012). This requires that education systems and institutions change and continue changing to keep up with the demands and opportunities afforded by new technologies and information.

The impact of these changes in education should be observed in the school and national curricula. For example ICT literacy has moved from being an additional skill to being embedded within the curriculum (Wilson, Scalise, & Gochyyev, 2015). In Australia, acknowledging the importance of student ICT skills in the workplace, the national curriculum views student ICT capability as built of five interconnected concepts: *Applying social and ethical protocols and practices when using ICT; Investigating with ICT; Creating with ICT; Communicating with ICT*; and *Managing and operating ICT* (ACARA, 2017). Each of these concepts addresses different aspect of ICT capability that should be integrated in all learning areas, ensuring a holistic approach to build student ICT capability.

These changes in societal expectations and school curricula raise issues surrounding the use of ICT in school (Ridgway, McCusker, & Pead, 2004). These issues have become more sophisticated than simply whether or not schools should allow students to access the Internet. The use of ICT in everyday teaching and learning activities in school needs more open access to internet connection, bringing with it different concerns. ICT-related concerns in school have now shifted into digital integration, such as the ways to teach cyber safety, online collaboration, and online research skills (Edwards et al., 2016; Moore, 2016). Schools are grappling with various resource implementation strategies such as Bring Your Own Device (BYOD) or one-to-one laptop (1:1) (Hynes & Younie, 2017; Keane & Keane, 2017). Schools are addressing the need to have policies regarding cyber safety, which include privacy, cyber bullying, and Internet addiction because these problems have been found to affect students' cognitive ability, mental health and well-being (Cross et al., 2016; Lindenberg, Schoenmaekers, Halasy, & Rehbein, 2017).

If schools are going to have ICT-rich learning environments, then they require relevant and suitable assessment methods (JISC, 2010; Masters, 2013). Authentic assessment in this kind of learning environment needs to employ more sophisticated tasks and measures of achievement supported by ICT. Unfortunately, with regards to the use of ICT, there is still a gap between learning activities and assessment, especially in high-stakes assessment (Miller, 2011; Timmis et al., 2016). While the aim has been to embed ICT in learning activities in school, with a priority to build student ICT capability; currently high-stakes assessment still relies mainly on paper-based activities and the traditional analytical marking method.

The rationale for the use of ICT in assessment is twofold. Firstly, because of the immersion of ICT in teaching and learning, assessment needs to incorporate ICT in its processes. Secondly, because of the advantages and availability of ICT in assessment, assessment processes should be digital (Masters, 2013). The potential to use ICT in educational assessment is immense and the advantages are many, such as for assessing problem solving skill, higher order thinking ability, and contextual understanding (Lin & Dwyer, 2006; Timmis et al., 2016). If the last decade of ICT development can be taken as a measure, the future should open up even more opportunities in the use of ICT as well as compel the integration of ICT in assessment.

To summarise, the change towards the digital era makes changes in educational policies a necessity. Changes in society lead to the need for different skill sets, which in turn leads to changes in schooling outcomes with ICT capability integrated in them (DiCerbo, Behrens, & Barber, 2014; Scardamalia, Bransford, Kozma, & Quellmalz, 2012). Educational practice and curricula have been changing to answer the new demands of the new direction. These changes call for a measurement method that is more suitable for the challenges presented in the new authentic form of assessment.

Rationale and Significance

Assessment of creative performance tasks such as design portfolios, live performances and artefacts has been problematic (Dorn, Madeja, & Sabol, 2004; Koretz, 1998; Wiggins, 1990). The variations of components in this type of student work, their subjective nature, the breadth and depth that characterise creative work, the richness of student experience that might set up the foundation of their ideas, and the abstractness of the thinking process that culminates in the production of the work, all contribute to the complexity of the judgement process. All these factors could result in inconsistencies in judgement (Traub & Rowley, 1991).

Compared to the simplicity of objective tasks such as multiple choice or short answer, for which the answer could easily be judged as right or wrong, judgement of the quality of creative production tasks could be considered a *wicked problem* (Henderson, 2014; Stables, 2017b), *ill-structured* (Anderson, 2016; Archbald & Newmann, 1988; Jonassen, 2003), or *messy* (Anderson; Wolf, 1989); a problem that does not only have a single solution. This kind of problem requires continuous effort and discussions to find the best ways to deliver assessment results that are accountable. The judgement result of complex practical tasks should highlight the richness of students' creativity and innovation instead of flattening the qualities into a one-dimensional score.

The importance of high-stakes assessments highlights the necessity for the responsible authorities to find methods to represent the quality of student work. High-stakes assessment results inform the decisions on whether a student advances to a higher level of study, passes a program, or obtains a certification (Kaufman, Graham, Picciano, Popham, & Wiley, 2014; Madaus & O'Dwyer, 1999; Ridgway et al., 2004). The student's future, the credibility of the education institution and the student's contribution in the community depend on the quality of the assessment result. Therefore, it is critical that assessment tasks can elicit performances that represent what is to be measured, and that judgements are accurate and reliable.

Traditionally, most high-stakes assessment in creative production is judged or scored based on a set of criteria in an analytical marking process (Taylor, 2006; Thorndike & Thorndike-Christ, 2010). This method should create a standard upon which the assessors base their judgements. The criteria should provide detailed descriptions that define the quality constituted by each score, hence offsetting variations in the judgements. In reality, however, many studies have indicated concerns over the reliability of analytical marking (Humphry & Heldsinger, 2014; Miller & Linn, 2000; Pollitt, 2004; Wiggins, 1990). An alternative judgement method that could produce assessment results that are more reliable, and therefore more accountable, is needed. One option is Comparative Pairs judgements.

The theoretical and statistical underpinning of the Comparative Pairs judgements method is Thurstone's Law of Comparative Judgement (Thurstone, 1927). In 1927 Thurstone proposed to measure psychological traits using comparisons. In education, his theory could be applied to the scoring of student works. In this method, instead of assigning a score to a student work, the assessors simply compare a pair of works and decide which one is better based on a holistic criterion. A statistical model based on Thurstone's work is then used to position those works on a scale, based on a logistic function of the wins and losses. While Thurstone proposed this theory almost a century ago, manual implementation had been impractical. This method required long and complex processes in both its judgements and analysis processes. Therefore it only became practical with the development of powerful software towards the end of the 1990s (e.g., RUMM) and later online tools (e.g., Adaptive Comparative Judgement system).

Based on Thurstone's Law of Comparative Judgement (Thurstone, 1927), Laming's proposition on human judgements (Laming, 2011) and the Rasch logistics model parallel to Thurstone's (Andrich, 1978; Pollitt & Whitehouse, 2012), the Comparative Pairs judgements method was designed. Recent advancement in computing technologies has made this method more feasible. A computer program can generate the pairing of student work, facilitate the judgement process, and generate the scores and reliability measures (Pollitt, 2012b). With both the student work and the marking interface digital and online, any number of assessors can judge the student work from anywhere in the world. The judgement sessions can continue on until a certain level of reliability is achieved. The software can then generate the scores based on the judgements. The results, the analysis of the results, and judgement processes data can subsequently be available for recording and analysis purposes. Over the past decade, the Comparative Pairs judgements method has been trialled with promising results in several countries and in various subjects for example Design (Kimbell et al., 2009), Engineering, Applied Information Technologies (AIT) (Newhouse, 2017), Mathematics (Jones, Swan, & Pollitt, 2015), and Chemistry (McMahon & Jones, 2015).

By investigating the suitability of the Comparative Pairs judgements method to assess two different creative tasks in high-stakes assessment, this study contributes to the understanding of this method and informs decisions on the use of this method of measurement. Consequently, this study examined the reliability of the scoring results, the validity of this assessment method, and the issues that arose from the implementation of this method of measurement. Findings from this study in turn contribute to the effort to find the assessment method that is most appropriate to assess students’ achievement in this information age.

Aims of Study

This study aimed to investigate the suitability of the Comparative Pairs judgements method for high-stakes creative production assessment. Comparative Pairs judgements method as an alternative method of scoring has been considered to have the potential for more holistic judgement of students’ learning than the traditional analytical method of assessment (Jones & Alcock, 2014; Pollitt, 2004). Several factors related to the suitability of this scoring method were explored. Such factors were the perception of the assessors and their response to the different scoring methods, the reliability of the scoring results and the validity of the assessment method, and the suitability of this method for different types of tasks in Design and Visual Arts.

Consequently, the study aimed to reveal characteristics of performances that are related to the quality of the digital representations of the performances, online scoring interfaces, and the judgement processes. The perception of assessors is crucial to success in using a method of scoring. Therefore, the study aimed to investigate the perceptions of assessors on the quality and limitations of the digital representations, the accessibility of the online scoring process, the method of scoring, and the quality of student work in general. Assessors' perceptions on these factors could highlight issues related to the digitisation and judgement processes, underline the factors that could affect the quality of the digital representations and the scoring results, and provide the information necessary for improvement.

The study considered different types of performance and compare the characteristics of performances as this was expected to lead to an understanding of the suitability of the Comparative Pairs judgements for different types of tasks. The importance of high-stakes assessment makes it critical that the rigor of the assessment method is examined within relevant boundaries that are the particular types of tasks. By comparing two different types of tasks, which were a finished product in Visual Arts and a process portfolio in Design, this study was expected to attest the suitability of the Comparative Pairs judgements method on similar types of tasks.

Problem Statement and Research Questions

The study was built on the concerns that in courses where creative production tasks are used, judgements are not comparable between contexts, are not reliable due to the subjectivity of assessors, and are not cost-effective for large groups of students spread across large jurisdictions. The Western Australian courses of Design and Visual Arts were chosen as illustration of these problems. For example, in the Visual Arts course in WA student *portfolios* (termed practical submissions) may include artistic artefacts that are two-dimensional, three-dimensional or motion and time-based. Thus judgements must be comparable between the different media of the artefacts, which highlights the issue of the adequacy of digital representation. Further, in the Design course detailed design documents are submitted to explain the development of design artefacts whereas in the Visual Arts course a very limited *artists statement* is submitted. Judgements must take account of differing amounts of accompanying information supporting the submitted artefacts.

The subjective nature of the tasks in the Design and Visual Arts courses is considered to be a source of concern. The Design task was a portfolio that contained various forms of evidence of the design process. The different types of evidence that the students chose to represent their design projects, the quality of their design process including innovativeness and creativity, and the students' ability to communicate their design could potentially result in variations of judgements among assessors. In the Visual Arts course, these variables were even more varied because the type of the assessment task was an artwork that could be in different forms, use a wide range of materials, show diverse techniques, or be comprised of a number of components. It becomes difficult to be sure what formed the basis of judgement and how consistent that would be between assessors and between student works.

Where the assessment is summative in nature (i.e., designed to determine the achievement of a student at the end of a learning sequence rather than inform the planning of that sequence for the student), it is critical that judgements reflect performance reliably (referring to the extent to which results are repeatable) and validly (referring to the extent to which the results measure the targeted learning outcomes). Drawing from two main fields of research: constructivism and educational measurement, the present study was focussed on the use and quality of the Comparative Pairs judgements method in a high-stakes assessment. The assessment tasks investigated were authentic performance tasks, as constructivist tasks largely are, and the task assessment was based on a Rasch model of measurement used in the Comparative Pairs judgements scoring. In doing so, it aimed to consider whether in the context of summative assessment the methods of judgement were manageable in terms of cost and particular characteristics of the assessment environments, which included the physical environment and digital environment such as Internet access.

Additionally, the need to manage, transport, and store student work creates logistic problems. The vastness of Western Australia exacerbates this problem. While transporting student work from metropolitan schools to the marking site might not be too difficult, it is not the case with country schools. With an area of 2,526 million km^2 (Australian Bureau of Statistics, 2017), transporting student work from country schools to the marking site in Perth could be challenging and costly, hence there needs to be an alternative way to assess creative expression work. The use of digital representations of student work in online scoring processes could be a good solution to this problem. With the digital representations of student work available online, there is no more need for the original student work to be transported and stored for scoring. However, efforts to be taken to ensure that issues such as the quality of the digital representations and the accessibility of both the digital artefacts and the online scoring systems were appropriate and therefore did not reduce the quality of the assessment results.

This study aimed to investigate the suitability of the Comparative Pairs judgements method in assessing digital representations of student practical task in high-stakes assessment, as exhibited by the validity and reliability of the scores resulting from the assessment process. Accordingly, this study was focussed on the research question:

How representative are the Comparative Pairs judgement scores of the quality of the student practical production work in Visual Arts and Design courses?

Within this research question, there were three subsidiary research questions that built into it:

In assessing student work in each of the Visual Arts and Design courses,

- *how valid and reliable are the scores and rankings generated by the Comparative Pairs judgements?*
- *what are the differences and similarities of the results from the Comparative Pairs judgements with the traditional analytical marking?*

- *how do the different types of work in Design and Visual Arts courses affect the scores and rankings generated by the Comparative Pairs judgements?*

Purpose of Study

The purpose of this study was to analyse the Comparative Pairs method of scoring as a method to assess students' practical creative work. The two secondary school courses chosen to test this method were Visual Arts and Design because both contained a prominent practical component beside a written component in their summative assessment. However, each had a different form of assessment: an artwork product and a design portfolio. By cross examining the Comparative Pairs judgements with the traditional Analytical marking for each course, this study aimed to build a better understanding of the reliability and validity of Comparative Pairs judging in different forms of practical work. By comparing the results and issues found in these two different courses, this study engaged in a discussion over the appropriateness of the use of Comparative Pairs judging in different school courses. With a better understanding of the Comparative Pairs judging in Visual Arts and Design courses, this study sought to contribute to the general knowledge of assessment, especially of practical production work, and specific knowledge on the use of the Comparative Pairs judgements method.

Scope and Context

This study was conducted within the first phase of an Australian Research Council (ARC) Linkage Project titled *Authentic Digital Representation of Creative Works in Education*, a collaborative project between the Centre for Schooling and Learning Technologies (CSaLT) of Edith Cowan University and the Curriculum Council, of Western Australia. This first phase took place in the first year of the four-year project. Two secondary school Western Australian Certificate of Education (WACE) courses, Design and Visual Arts, were investigated in both the present study and the main project.

The main project was a four-year project that was divided into three phases. In the first phase, *Development & Pilot*, the researchers including the author digitised the student practical work in Design and Visual Arts that was submitted to the Curriculum Council for the WACE examination. Two online scoring processes, the Comparative Pairs judgements and the traditional Analytical marking, used this digital representation of student work. Scoring data from these online processes and the WACE practical examination, together with interviews with the assessors, teachers and students, were collected. Because the present study was focussed only on the suitability of the Comparative Pairs judgements, only the scoring data and the interviews with the assessors, and assessor notes from the judgements process, were analysed. A detailed discussion of the main project and the way this study was embedded in it is presented in Chapter 3.

Limitations of Study

This study investigated the suitability of the Comparative Pairs judgement for assessing practical tasks in high-stakes assessment. Two secondary school courses with different types of practical task were chosen: Design and Visual Arts. This study is limited to the examination of data from the scoring methods and assessor interviews, therefore there was no in-depth discussion on subject-related possible factors such as factor analysis on the analytical rubric and subject-related pedagogy theories.

In this study no data from student surveys and interviews was used because of the limitation on the scope of the study. However, it was recognised that the student could be an important factor that affected the validity of the scoring results, in particular student confidence in presenting their work would be a factor. In the Design course at least the work showed how the students progressed, in VA it was only one artwork that was supposed to represent the whole achievement in learning. However, these factors were not directly relevant to the method of judging.

In this study, the term *validity* is discussed in the more narrow term of *suitability*, which refers only to the suitability of the Comparative Pairs judgements method as an alternative method for assessing practical tasks in high-stakes assessment. This more limited term is meant to provide a perimeter for the discussion of the validity of the scoring method so it does not move beyond this boundary into a broader definition of validity such as face validity, construct validity and so on.

For the Comparative Pairs judgements online assessment system, this study utilised the Adaptive Comparative Judgement system developed by the TAG Learning company (Pollitt, 2012a). As such, this study took assumptions that the system algorithm was fully tested by the developer and was a valid system. The system has been tested in large projects in a number of places such as the United Kingdom, the United States, and Sweden (Bartholomew, Hartell, & Strimel, 2017). This study recognises, however, that the online digital system could have weaknesses.

Educational assessment should be driven by pedagogy and thus focussed on student learning. However, in this study the pedagogy that drives the assessment is only acknowledged as the broader background and not discussed in detail because of the limitations of the study and also because the focus of this study was on summative assessment, which has the main purpose to provide summarised information for higher education institutions or workplaces regarding student achievement at the end of secondary schooling. Therefore it is not designed to specifically inform pedagogy.

Structure of Thesis

This thesis is organised into seven chapters. Chapter 2 presents a review of the literature that underpinned the structure and nature of the study. It begins with a theoretical framework and finishes with the conceptual framework. Chapter 3 outlines the method used in the study to address the research question. This methodology chapter is followed by two data analysis chapters: data analysis from the Design course in Chapter 4, Visual Arts in Chapter 5. The cross-case analysis of Design and Visual Arts is presented in Chapter 6 and is based on the conceptual framework, followed by a discussion of the findings in terms of the research questions. Chapter 7 concludes this thesis with a discussion of the findings in terms of the research questions, and discussions of implications for policy, practice, and further research.

CHAPTER 2
LITERATURE REVIEW

This chapter discusses a review of the literature associated with the theory and research upon which this study was based. The review starts with the theoretical framework for this study and a discussion on assessment, especially on digital assessment and authentic assessment. This is followed with a discussion on the components of assessment, particularly those related to scoring and on assessment qualities pertained to this study. The theoretical framework presents the epistemology that underpins the main theoretical components of this chapter to define the structure of this study. The theoretical components related to this framework are discussed in the sections that follow. This chapter concludes with a conceptual framework for the present study, which portrays the relationships among these theoretical components and the way this study fits within these related theories.

Theoretical Framework

The theoretical framework that grounded this study was built upon two areas of theory: constructivism; and a theory of educational measurement, the Rasch models in particular. Constructivism provided the rationale behind the focus on authentic assessment and assessor perceptions; while the educational measurement theory provided the vehicle for assessment methods that could appropriately be applied to the type of performance assessment that was the focus of this study. The relationships between these areas of theory as related to the focus of this study are presented in *Figure 2.1*.

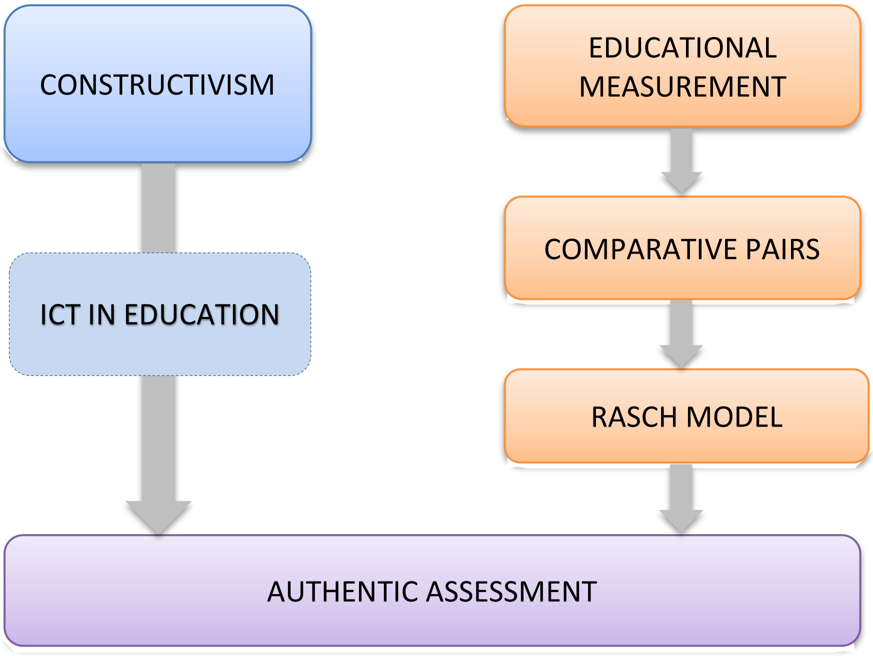


Figure 2.1 Theoretical Framework

Constructivism

The theory of constructivism views learning as a personal process in which learners construct new understanding based on their existing knowledge and personal experience through activities that are contextual, relevant, and meaningful (Driscoll, 1993; Ertmer & Newby, 1993; Karagiorgi & Symeou, 2005; Vygotsky, 1980; Wiggins, 2015). A theory built upon the contributions of psychologists, educators, and researchers such as Vygotsky, Piaget, Bruner, Papert, Jonasson, Wiggins, and many more over the years; constructivism is supported by theorists in various fields such as biology, psychology, pedagogy, linguistics, and neuroscience (Wiggins, 2015). While constructivism is often viewed as a learning theory, in a broader sense it is an epistemology, i.e., a theory of knowledge, that has been implemented in learning activities in both formal and non-formal settings (Jonassen, 2006a).

As an epistemology, constructivism does not stand alone (Driscoll, 1993). Instead, it originates and employs aspects of various learning theories such as behaviourism, cognitivism, and neuroscience. Jonassen (2006a) advised that "…view that objectivism is antithetical to constructivism is an impoverished view of constructivism. … The reality is that very little of the knowledge that learners construct can be predicted by any model of instruction or theory of learning" (p. 44). Constructivism could take advantage of behaviourism since it views learning as building new knowledge based on previous understanding and experience. Even though behaviourism is more outcome-based while constructivism is learner-based (Ferguson, 2001), certain behaviourist approaches such as concept generalisation and reinforcement (Ertmer & Newby, 1993) could be utilised within constructivist learning environment to help students build association with new constructs. It makes use of cognitivism and neuroscience in helping students code and construct new knowledge to make it meaningful and permanent (Mareschal, Butterworth, & Tolmie, 2014). In essence, constructivism views learning as a holistic, authentic, student-centred, multimodal and personal learning experience. This clearly aligns with the use of digital technologies to support authentic assessment.

A constructivist epistemology considers that the construction of new knowledge is situated in action and context, otherwise known as *situated learning* (Brown, Collins, & Duguid, 1989; Driscoll, 1993; Merrill, 1992). Therefore, learning activities built on a constructivist view typically employ diverse methods, media, and situations to help students construct their knowledge in a contextual and meaningful real-life situation (Ertmer & Newby, 1993; Wiggins, 2015). Constructivist learning environments provide students with the opportunity to collaborate, to think, to solve problems, to create, and to evaluate their own learning (Begg, 2015; Binkley et al., 2012). Tasks are implemented that are designed to stimulate students' ability to make meaning out of their activities and scaffold their understanding and skills as they progress. This characteristic of constructivist learning process could be well facilitated with ICT (Anderson, 2016; Dunleavy & Dede, 2014; Newhouse, Trinidad, & Clarkson, 2002; Perkins, 1992b).

ICT in Education

ICT has been considered essential to bridging the gap between the confines of the physical world and the complexity of learning environments needed in constructivism (Allison & Kendrick, 2015; Perkins, 1992a; Spector, 2014). Digital technologies could create learning environments that are otherwise challenging or even impossible in reality with facilities such as online communication, computer simulations, and design applications. Students can use computer simulations and games to engage in an immersive learning experience that is otherwise too difficult, too expensive, or dangerous to do, for example in astronomy, physics, human biology, history, and chemistry (Andrews & Wulfeck II, 2014; Dawley & Dede, 2014; Masek, Murcia, Morrison, Newhouse, & Hackling, 2012; Steinkuehler, Squire, & Barab, 2012). Online communication, social media, and learning management systems such as Firefly, Google Classroom, and Moodle make collaboration and communication among students, teachers, and parents convenient (Henderson, Snyder, & Beale, 2013; Hsu, Ching, & Grabowski, 2014; Wilson, 2017). Numerous software and apps are available to learn design and creation, for example audio and video editors such as iMovie, graphic editors such as Adobe Photoshop, and design modelling and manufacturing software such as Autodesk Revit (Bruce & Chiu, 2015; Henderson et al., 2010; Roads, 2015).

Parallel to this, development in ICT has opened up new possibilities for teachers to facilitate student learning with tasks and learning environment that are situational, contextual, and meaningful (Howland, Jonassen, & Marra, 2012; Jonassen, 2006b). Howland et al. (2012, p. 3) identified five attributes of meaningful learning, which articulated meaningful learning as active, constructive, intentional, authentic, and cooperative. These five attributes could be well supported by the use of ICT in student learning, as Howland et al. further attested. Among these functions of ICT are to facilitate research and collaboration, to simulate situations to provide contextual learning, as a tool to create and compose, to record student learning and achievement, and many more (JISC, 2010; Lockee & Wang, 2014). The availability of many ICT resources could help teachers to provide a learning environment that is better aligned with constructivism. However, planning and implementing such programme is difficult and requires skills and experience.

The interaction among teaching components in a technology-rich student-centred classroom is complex, particularly when including assessment. Teachers’ role in this kind of classroom is demanding and challenging. A framework that is commonly used to prepare teachers to integrate technology in this classroom is the Technological Pedagogical Content Knowledge framework (TPACK) (Koehler, Mishra, & Cain, 2013; Mishra & Koehler, 2006) as shown in *Figure 2.2*. The TPACK framework outlines the interaction among the components of technology integration in teaching with the "complex interplay of three primary forms of knowledge: Content (CK), Pedagogy (PK), and Technology (TK)" (Koehler, 2017) at the centre of the interaction. As Koehler et al. (2013) explicated,

> TPACK is the basis of effective teaching with technology, requiring an understanding of the representation of concepts using technologies, pedagogical techniques that use technologies in constructive ways to teach content, knowledge of what makes concepts difficult or easy to learn and how technology can help redress some of the problems that students face, knowledge of students' prior knowledge and theories of epistemology, and knowledge of how technologies can be used to build on existing knowledge to develop new epistemologies or strengthen old ones. (p. 16)

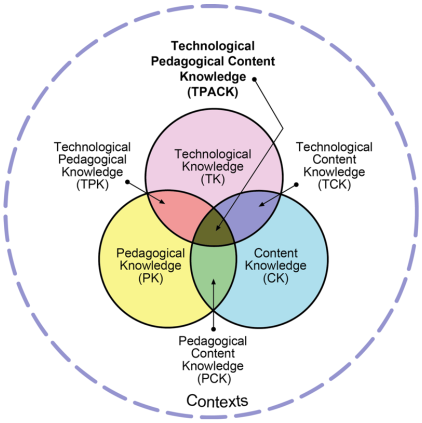

Technological Pedagogical Content Knowledge (TPACK)
Technological Pedagogical Knowledge (TPK)
Technological Knowledge (TK)
Technological Content Knowledge (TCK)
Pedagogical Knowledge (PK)
Content Knowledge (CK)
Pedagogical Content Knowledge (PCK)
Contexts

Figure 2.2 TPACK Framework (Koehler, 2017)

ICT integration in teaching and learning provides a support for constructivist learning environments that encompass curriculum, pedagogy, and assessment (Lim et al., 2013). Authentic, contextual learning embodied in constructivist learning environment requires authentic assessment (Howland et al., 2012). In the past, educators found conducting authentic assessment difficult because it is usually difficult to conduct and measure because this kind of assessment is usually multimodal and time-sensitive (Scardamalia et al., 2012; Wiggins, 1990). The contrast between the limitations of the traditional pen-and-paper assessment and the appropriateness and availability of ICT in facilitating more authentic forms of assessment, makes a case for the use of digital assessment.

The term *authentic assessment* was initiated by Wiggins in 1989 (Wiggins, 2011) when he proposed what defined a true test: "… we have lost sight of the fact that a true test of intellectual ability requires the performance of exemplary tasks" (p. 81). He continued on defining authentic assessment as "not only reveals student achievement to the examiner, but also reveals to the test-taker the actual challenges and standards of the field" (p. 82). In essence, authentic assessment refers to assessment that measures students' contextual knowledge and skills. In practice, authentic assessment is characterised as meaningful; showcasing students' mastery; scaffolding students' competence and higher-order thinking ability; contextual; situational; viewing learning as a process in which "'content' is to be mastered as a means, not an end" (Wiggins, 2011, p. 91); and including collaboration, feedback, problem solving, and synthesis (Archbald & Newmann, 1988; Ashford-Rowe, Herrington, & Brown, 2014; Lund, 1997; Wiggins, 1990, 2011).

Parallel to the use of ICT in constructivist learning environments, the use of ICT in authentic assessment has created the opportunity to create assessment processes that fits its purpose. Digital assessment could be designed to assess students' authentic learning in different subjects for different purposes through audio recording, video recording, Computer Adaptive Test (CAT), e-portfolio, simulation games, and many more (Clarke-Midura & Dede, 2010; Crisp, 2009; Howland et al., 2012; Kimbell et al., 2009; Madeja, Dorn, & Sabol, 2004; Newhouse et al., 2011; Williams, 2009). ICT has afforded education with means to facilitate various assessment tasks and the assessment of these tasks. The type of task that was suitable for the purpose of the assessment and the method with which the task was assessed should be considered with care to ensure the quality of the assessment result.

The subjectivity and individual nature of authentic assessment often makes ensuring the reliability and validity of the assessment results challenging. Typically assessors need to make highly subjective judgements. Human judgement tends to be limited, subjective, and unreliable (Laming, 2011). Moreover, parallel to the way students construct their understanding through previous personal experience, assessors also construct their judgements through their experience in their fields (Bloxham, den-Outer, Hudson, & Price, 2016). In their judgements, assessors draw from their own constructed view on how the learning results should be, what should be assessed and what the standards are. This personal view has been known in different terms such as “standards frameworks”, “teachers conceptions of quality … [and]… interpretive frameworks” (p. 13). Research suggests that even with the use of marking schemes such as analytical rubrics this factor is still influential, especially in fields that are subjective and creative such as arts (Dorn et al., 2004; Humphry & Heldsinger, 2009; Pollitt, 2004), because assessors could still be biased towards varied qualities.

Research has shown that the use of good quality marking schemes could increase the reliability of scores and reduce bias (Andrade, 2005; Jonsson & Svingby, 2007; Moskal & Leydens, 2000). Marking schemes are considered of good quality if they are well designed and provide clear descriptions of outcomes. However, caution needs to be taken because even if a marking scheme proves to have a high reliability, it does not necessarily indicate high validity, as Jonsson and Svingby further explicated “Just by providing a rubric there is no evidence for content representativeness, fidelity of scoring structure to the construct domain or generalizability” (p. 137). Furthermore, marking schemes are designed with the main purpose to increase the reliability of scores; therefore they usually contain measurable outcomes (Popham, 1997). However, in authentic assessment not all outcomes are easily measured, and attempts to impose measurable outcomes on a marking scheme could compromise the validity of the assessment for the sake of reliability (Whitehouse & Pollitt, 2012). Andrade (2005) suggested that for marking schemes to be valid, not only do they need to address reliability, they also need to address equity and alignment to standards and the curriculum.

The complexity and depth of authentic assessment tasks call for scoring methods that are beyond an aggregate grade that is based on a set on predetermined criteria (Wiggins, 2011, p. 91). Currently, the most common assessment method is the analytical marking with an analytical marking rubric (Jonassen, 2014). This scoring method was considered to provide scoring results that were reliable and valid. However, concerns over this claim and current development in digital technologies signalled the need for a better method, one that was more suitable for the characteristics of authentic assessment. The Comparative Pairs judgements method has been considered to be such method (Kimbell, 2007; Pollitt, 2012b). This judgements method is based on educational measurement theory represented in Rasch models.

Educational Measurement and Rasch Models

Educational measurement is based on a premise within psychology measurement, or psychometry, that psychological traits such as student attainment could be measured quite similarly to the measurement of physical properties (Bond & Fox, 2001; Fischer & Molenaar, 1995; Thurstone, 1928). Theories related to educational measurement have been progressing from a simple pass-or-fail to statistical model fitness currently used. One of the family of statistical models commonly used today is the Rasch method, mostly because it can position the estimates of both student ability and item difficulty parameters in one scale (Bond & Fox, 2001). This section presents a brief discussion on educational measurement which leads on to the rationale for the use of Rasch models.

Psychometry or psychological measurement is the measurement of psychological traits. In psychometry, psychological traits such as happiness, violence tendency, and achievement are assigned a numerical value in a measurement scale, which is not too different to the measurement of physical attributes such as mass and length, albeit more complicated and less consistent (Pellegrino, Chudowsky, & Glaser, 2001; Raykov & Marcoulides, 2011; Thurstone, 1928). According to Pellegrino et al. (2001), educational assessment uses psychometric models that “are based on a probabilistic approach to reasoning” (p. 112). They went on to describe that, “a statistical model is developed to characterize the patterns believed most likely to emerge in the data for students at varying levels of competence" (p. 112).

Historically, the impetus of psychometry began in late 1800s in experimental psychology followed by the use of statistical analysis in this field (Ward, Stoker, & Murray-Ward, 1996). Later in early 1900s psychometry started to be used in education and further development of psychometry theory reached its culmination between 1930 and 1960. In most part of this period, the use of the Classical Test Theory (CTT), which was also called the Traditional Test Theory (TTT), was more prominent until Dr. Frederic Lord and Georg Rasch proposed the Item Response Theory (IRT) and the Rasch model, respectively. These were developed to better represent student scores because unlike CTT, IRT considered test difficulty and variations of score distributions as influential in predicting student ability (Andrich, 2004; De Ayala, 2009; Fischer & Molenaar, 1995; Ward et al., 1996; Yu, 2011).

Unlike CTT, both IRT and the Rasch model calculate these variables and provide a probabilistic analysis of both item difficulty and person ability parameters. The more simple variation of the Rasch model, which is called the Rasch dichotomous model, is used to analyse data that only consist of two responses such as True or False questions, or as in this study, win or lose as judged by the assessors. This analysis is similar to a special case in IRT and is sometimes referred as such, even though both models were developed separately (Andrich, 2004; Embretson & Reise, 2013). In this study, the Rasch dichotomous model was used, and therefore will be discussed later in this chapter. The Rasch dichotomous model is parallel to Thurstone's Law of Comparative Judgement (Andrich, 1978; Thurstone, 1927) and served as the base for the Adaptive Comparative Judgement (ACJ) system used in the Comparative Pairs judgements method in this study. For test items with more than two ordered responses, for example Likert scale, the more general Rasch polytomous model is used (Andrich, 1978).

The theoretical framework of this study was based on a constructivist view of learning that lends itself to authentic assessment relying on assessor judgement, and drawing on the Rasch model as the appropriate measurement method. The following sections discuss further the components within these theories that are pertinent to this study.

Educational Assessment

Assessment is an integral part of education. It provides the information that is essential for the next steps in the teaching and learning process. It serves as the foundation upon which improvement is planned and implemented. It has an impact on student learning and direction for future studies or employment. It has a determining role in the educational policy-making process across the local school, state, national, and international settings. This section discusses a general overview of educational assessment with a focus on digital assessment.

In general, the term *educational assessment* refers to the "procedure for eliciting evidence that can assist in educational decision-making" (Wiliam, 1994, p. 5). This definition refers to the entire process of assessment from designing, conducting, recording, scoring, decision-making, and reporting assessment of student learning. Methods to assess student learning and achievement could vary from as simple as observing primary school students writing a word, to a nation-wide standardised secondary school examination (Dietel, Herman, & Knuth, 1991; Gipps, 1994). Assessment tasks could be written, performed, typed, spoken or created multimodal. Taras (2005) defined the process of assessment as "the mechanics or steps required to effectuate a judgement … [and that] … A judgement cannot be made within a vacuum, therefore points of comparison, i.e. standards and goals, are necessary" (p. 467). The process of assessing student work or performance is fundamentally a process of comparing the work or performance to a standard or to another student's work. Throughout this thesis the word *judgement* is used to express the process of assessing student work across different methods.

The main purpose of educational assessment is to provide information to make a decision (JISC, 2010; Miller, Linn, Gronlund, & Linn, 2009). This decision could be a remedial, an award, an inclusion or others. It could be communicated with students, parents, school, or it could simply remain a record that provides diagnostic information for teachers. Table

2.1 delineates the classification of educational assessment based on the form of the assessment, the purpose of the assessment, and the method of interpreting the result; as adapted from Miller et al.'s classification (2009, p. 43). Based on its form, assessment could be classified as written, such as multiple-choice test, and performance, such as dance performance. Based on its purpose, assessment could be could be used for placement, formative, diagnostic, or summative purposes. The method used to interpret the assessment result could be criterion-referenced, in which achievement is assessed based on a set of criteria; or norm-referenced, in which student's achievement is ranked based on the achievement of his peer (Bond, 1996).

Table 2.1
Classification of Educational Assessment (adapted from Miller et al., 2009, p. 43)

Basis for classification	Type of Assessment	Function	Instruments
Form of assessment	Written test	Efficient measurement of knowledge and skills	Multiple choice test Essay
	Performance test	Assessment of performance of practical skills or problem-solving abilities	Laboratory experiment Dance performance Creation of an artwork
Purpose of assessment	Placement	Determines prerequisite skills and mastery for placement	Readiness test Aptitude test
	Formative	Determines learning progress and, when necessary, appropriate remedial lessons	Teacher-made test Classroom observation
	Diagnostic	Determines causes of persistent learning problems	Diagnostic test Observation
	Summative	Determines level of achievement at the end of a learning period	Standardised or teacher-made summative test
Method of interpreting result	Criterion referenced	Measurement of student achievement based on a set of criteria	
	Norm referenced	Measurement of student achievement based on the relative position in the cohort	

Based on the purpose of assessment, Scriven (1967) differentiated two types that were most common: formative and summative. However, even though there are certain differences in these two types, one could still serve the purpose of the other (Bennett, 2011; Harlen, 2005; Isaacs, 2013; Taras, 2005; Wiliam, 2000), and the two types could be viewed as two ends of a continuum (Stables, 2015). Originally, formative assessments sought to gather information regarding student achievement to plan for the next step of their learning. Because the focus of formative assessment is on using the result for students' further learning, it is also called the *assessment for learning* (Bennett, 2011; Isaacs, 2013; The Cross Sectoral Assessment Working Party, 2011). Formative assessments could take many forms, including questioning, classroom observation, quiz, short essay, reflective journal, and peer assessment (Bennett, 2011). In comparison, summative assessments are designed to measure the student achievement at the end of a program, for example a final visual arts project or a term examination. While the purpose of assessment may be summative, the assessment data could also be used for formative purposes, and vice versa. When applicable, data from summative assessment could be used to support learning while data from formative assessment could be used to summarise teaching and learning or suggest program improvement.

Summative assessment is usually conducted at the end of a learning period to measure student achievement, therefore it is also called the *assessment of learning*. This type of assessment is also sometimes referred to as *high-stakes assessment* because of its purpose as the basis to make highly consequential decisions (Bennett, 2011; Gunzenhauser, 2003; Isaacs, 2013; The Cross Sectoral Assessment Working Party, 2011). The conflict of the educational and social purposes of this assessment resulted in the negative connotation of the term *high-stakes assessment* (Kohn, 2000; Nichols & Berliner, 2007; Orfield & Kornhaber, 2001; Taras, 2005). The term *high-stakes assessment* and *summative assessment* are used interchangeably throughout the present study.

In Western Australia, the Western Australia Certificate of Education (WACE) examination, which is conducted at the end of year 12 to conclude the senior secondary schooling (Curriculum Council of Western Australia, 2011d), is a critical summative assessment. In general, the WACE examination results have both diagnostic and summative purposes. The results could provide the students with information on their academic achievement;

assist tertiary education providers and industry in entrance selection process; and contribute to the evaluation of teachers, schools and the education department learning and teaching programs (p. 63). The WACE examinations were the context for the present study.

Processes of Assessment

Assessment starts with the end, by first considering the objectives the assessment aims to achieve. Messick (1994) defined this process as construct-centered assessment proces:

> A construct-centered approach would begin by asking what complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. … Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. (p. 16)

In construct-centered assessment processes, the objectives are the driving force in the conception of the assessment task and the scoring method. The advantage of this process, as Messick (1994) further explicated, is that by using the construct to guide the whole assessment process, both the processes of task development and scoring are kept aligned to the objectives of the assessment, and thus ensuring the reliability of the scoring result and the validity of the assessment.

In general, the development of assessment processes could be seen as being related to two main components: the *assessment task* and the scoring method, or the *task assessment* (Campbell, 2008). These two components depend on the purpose of the assessment; i.e., placement, formative, diagnostic, or summative; and the constructs the assessment aims to measure. The assessment task is then developed with tasks that could be in the form of a written task, a performance task, or a combination of both. The development of task assessment follows next, by considering the method to interpret the result and the scoring method and the scoring schemes suitable for the method and constructs measured (Messick, 1994; Miller et al., 2009).

In the present study, the purpose of the assessment was summative. The form of the assessment task of this summative assessment was performance-based authentic assessment. The scoring methods within the task assessment used were the traditional analytical marking and the alternative Comparative Pairs judgements method. The scoring processes were conducted online using the digital representations of student work. The following sections discuss these components, starting with the concept of digital assessment to frame the discussion on assessment task and task assessment.

Digital Assessment

The 21st century has brought with it a new technology-rich landscape in education. In terms of teaching and learning activities, technology has enabled richer learning environments and resources as well as created the necessity for schools to teach students new sets of skills that could help them prepare for the challenges of the future (Binkley et al., 2012; DiCerbo et al., 2014; Griffin et al., 2012). Parallel to this, digital technologies have enabled and at the same time required new models of assessment. The process of creating tasks; facilitating, collecting and assessing student work; recording student achievement; creating feedback; and reporting to parents has become more flexible and feasible by the availability of ICT infrastructures in schools and homes (JISC, 2010; Stacey & Wiliam, 2012). On the other side, digital technologies have also required that students learn new skill sets, and these new skill sets necessitate new models of assessment, ones that are more relevant with the ICT-rich curriculum (Masters, 2013).

In ICT-rich curriculum, digital technologies use is inseparable from teaching and learning activities in school. Students and teachers interact in both physical and virtual environments through the use of Learning Management Systems (LMS), educational social media such as Edmodo and Moodle, and other multimedia platforms such as YouTube and blogs. Lessons are recorded, distributed and learned with the help of computer and other electronic devices. Student work could be created and saved digitally. These digital forms of student work could be in many variations such as word documents, spreadsheets, powerpoint presentations, blogs, videos, vlogs, eportfolios, and many more (Howell, 2013; McLoughlin & Lee, 2007; Selinger & Kaye, 2005). School has the responsibility to build student ICT capabilities to ensure no student is disadvantaged in this learning environment. The Australian Curriculum defines student ICT capabilities as (ACARA, 2017)

> …students develop Information and Communication Technology (ICT) capability as they learn to use ICT effectively and appropriately to access, create and communicate information and ideas, solve problems and work collaboratively in all learning areas at school and in their lives beyond school. ICT capability involves students learning to make the most of the digital technologies available to them, adapting to new ways of doing things as technologies evolve and limiting the risks to themselves and others in a digital environment. (para. 4)

The digital form of student work has many advantages compared to the physical form. Digital work could be backed up, copied and sent to different recipients in a matter of seconds, thus reducing problems associated with distance, space and time (Binkley et al., 2012; Newhouse, 2014). Digital technologies have made possible assessment activities such as marking by several assessors in different places simultaneously, online moderation, digital annotation on student work, reporting to parents and school record updates. This also makes it possible for teachers from remote schools to participate in online marking and online moderation, and for students from these schools to send their work for online examination and participate in online assessments such as OLNA (Online Literacy and Numeracy Assessment) (School Curriculum and Standards Authority, 2014a) and online NAPLAN (National Assessment Program – Literacy and Numeracy) (NAPLAN, 2013).

In practical tasks such as sport performance, musical performance, artwork creation and design development, in particular, the flexibility of student digital work is valuable. In an Engineering course for example, Newhouse (2011b) reported on a project in which the students designed and built their water purification system, and recorded the complete process in videos, notes and diagrams in an online assessment management system. The same project also captured the assessment of students who studied Italian, Applied Information Technology and Physical Education in various multimedia repositories. Findings from that project included students' satisfaction and that

> …opportunity to demonstrate their creative capability in examinations situations in courses such as AIT and Engineering…. Digital assessment provides the ability to capture student knowledge and performance using a number of media (text, images, sound, video, etc) and this provides an improved and more authentic method compared with the traditional paper and pen method of assessment. (p. 16)

Digital assessment has been investigated in many places around the world across different subjects. In the Creative Arts, Dillon and Brown (2006) examined and later advocated the use of ePortfolio, "to manage media-rich expressions and representations of human activity in an integrated fashion" (p. 419). In Science, Geography, Design and Technology, Engineering, and Italian, an online assessment management system called eScape was tried and found successful in capturing, collating, and scoring students' practical performance and artefacts (Kimbell et al., 2009; Newhouse et al., 2011). In mathematics Stacey and Wiliam (2012) reported the flexibility and accuracy of digital assessment. Most findings on these studies indicated the benefits of digital assessment, including the affordability, flexibility, and the reliability of digital assessment. Similar to the traditional assessment, the quality of digital assessment depends on the assessment task and the task assessment. The following sections discuss these two components of assessment.

Assessment Task

The present study views assessment as comprised of two intertwining sides, which are the student side, or the Assessment Task; and the assessor side, or the Task Assessment (Campbell, 2008, p. 23). The Assessment Task side pertains to stages such as the planning, the development, the pilot testing, the administration, and the submission of the task. It therefore could be affected by factors that arise from those stages such as the quality of the task, the feasibility and the manageability of the task administration, and the alignment between the type of the task with the purpose of the assessment.

As was previously discussed, construct-centered assessment processes commence with the setting up the assessment objectives, followed by the development of the task and the scoring method. Consequently, the quality of the assessment result is influenced by the quality of each assessment process and the way all elements support one another. From the assessment task side, there are matters concerning the quality of the task such as the extent to which the task is related to the scoring criteria and the extent to which the task represents the knowledge or performance the assessment aims to measure, which are the elements of construct validity (Raykov & Marcoulides, 2011).

The discussion of assessment task in this section is limited to the performance-based authentic assessment task, as this form of task is the one investigated in this study. The relevance of this form of task with the nature of the Design and Visual Arts course is also presented. The reliability and validity measures of the assessment result are discussed later in this chapter.

Performance-based authentic assessment

Reservations about the quality and consequences of traditional objective assessment have sparked the popularity of performance-based assessment in the late 1970s. While the traditional assessment methods typically based on recall and writing were considered to restrict student learning experiences, performance-based assessment was considered to be more authentic and could provide more exhaustive data on student achievement and learning process (Hart, 1994; Kimbell et al., 2009). Performance-based assessments stimulate more complex thinking processes because they do not limit the responses to a set of choices or short answers (Jonsson & Svingby, 2007). Basically, this type of assessment refers to assessments that are considered to be more practical and more relevant to the real world, which should apply to all school subjects such as physical education, arts, mathematics and science.

The terms *authentic assessment*, *alternative assessment, performance assessment* and others are often used interchangeably depending on the emphasis of the discussion (Miller et al., 2009, p. 8). For example, the term *authentic assessment* is typically used to refer to the nature of an assessment that is designed to better assess the skills relevant to the real world by using more genuine evidence of student achievement. Archbald and

Newmann (1988) considered authentic tasks as “worthwhile, significant, and meaningful” (p. 10). Dorn, Madeja and Sabol (2004) defined authentic assessment based on the role of *factual knowledge* in the assessment by explicating that “Authentic assessment does not focus on factual knowledge as an end in itself. Rather, it focuses on the ability to use relevant knowledge, skills, and processes for solving open-ended problems during meaningful tasks” (p. 15). Wiggins (1990) made a comparison between authentic tasks and the traditional ones based on the directness of the learning evidence. Wiggins asserted “Assessment is authentic when we directly examine student performance on worthy intellectual tasks” (p. 2), while traditional assessment “relies on indirect or proxy ‘items’ – efficient, simplistic substitutes from which we think valid inferences can be made about the student’s performance at those valued challenges” (p. 2).

The term *alternative assessment* is used to highlight the difference between this type of assessment and the traditional pen-and-paper assessment (Andrews & Wulfeck II, 2014), which was usually in the form of multiple choice or short answer. The term *performance assessment* highlights the practical nature of the assessment, such as dance performances, science experiments, design portfolios or sports skills, (Beattie, 1997; Dorn et al., 2004; Miller et al., 2009). In the present study, the tasks that were assessed were in the forms of artwork in Visual Arts and product development portfolio in Design, therefore generally the term *practical production* is used. Even though technically the Visual Arts artworks are a finished product and the Design portfolios are an evidence of a process, they would be considered broadly as forms of performance assessment that are highly authentic.

Even though in many cases authentic assessment is considered to provide a more valid outcome, this type of assessment is also usually labour-intensive, time consuming and difficult to conduct (Wiggins, 1990). As time progressed since the dawn of authentic assessment, and with it advances in Information and Communications Technology, the ease with which this type of assessment could be conducted improved considerably (Brown & Dillon, 2006; Kimbell et al., 2009; Newhouse, 2011b; Newhouse & Tarricone, 2014). Current ICT development has made it possible for teachers and schools to have an electronic repository of student work in the forms of texts, drawings, photographs, videos, blogs and many others. It also has made possible the online assessment of student work, including enabling teachers from different locations to access and mark student work, and analyse the assessment results remotely. These technologies facilitate various uses in education and education assessment, such as Computer Adaptive Test (CAT) which presents successive questions based on students' ability as indicated by their response on the previous question and the Adaptive Comparison Judgement (ACJ) system used in this study, which creates pairing of student works, collects assessors' judgements and analyses the judgement data.

While the advancement of ICT in education has opened doors to various ways to record, assess and report student achievement, each of those ways came with unique challenges and limitations, especially on the constructs of validity and reliability (Dermo, 2009; DiCerbo et al., 2014; JISC, 2010). Educators and test developers always face the decision to choose the appropriate assessment as well as the best method to analyse and interpret the results, especially in high-stakes assessment. Factors to consider included human and technology resources, assessment systems, and psychometry.

When authentic assessments are in the form of a performance, evidence of a process, or a product, they require subjective judgements. This subjective nature could threaten the reliability of the score (Dorn et al., 2004; Koretz, 1998; Miller & Linn, 2000), as Traub and Rowley (1991) stated "…more reliable scores come from tests in which the items can be scored objectively than from tests in which the scoring involves an element of subjectivity" (p. 43). While in objective tests such as multiple choice or short answer there is little or no difference between different assessors, in a performance assessment, different assessors could potentially give very different scores for each assessed skill. When the total score is calculated, these differences are also added up, creating a larger difference in the total score even when a good scoring rubric is utilised (Jonsson & Svingby, 2007; Pollitt, 2012c).

Since it is imperative that school courses provide students with the knowledge and skills that are relevant to the real world, it follows that it is also important that the assessment suitably represents those knowledge and skills. The search for the best assessment methods for assessing authentic tasks has opened up discussions among educators, especially in courses with a major practical component. The varied issues that each of those courses have depend greatly on the nature of each course and the kind of knowledge and skills each course aims to develop. In this study, two senior secondary school courses, Design and Visual Arts, were investigated. While these courses were similarly practical and creative in nature, they have different issues and challenges, especially because of the difference in the type of the WACE examination tasks in the two courses. The next two sections present the philosophical and situational background on each course.

Design course

Design-related courses have been implemented in various forms and under different names in different countries, depending on the national educational focus (Banks & Williams, 2013). In the United Kingdom, for example, Design had mostly been taught in both a stand-alone Design subject and within Design and Technology, while in the United States of America, the general term is Technology Education. In Western Australia, beside the Design course that was investigated in this study, there was a separate course of Materials Design and Technology. While the Design course was focussed on different Design contexts which were photography, graphics, dimensional design and technical graphics (Curriculum Council of Western Australia, 2010a), the Material Design and Technology course was focussed more on specific design materials such as wood, metal and textiles. Kimbell (2011) posited that even if the Design term was dropped from the Technology course, it remained a Design and Technology course, parallel to Williams' assertion that design is inseparable to technology and technology education (2000). The current Australian national curriculum uses the term Technologies as an umbrella term for courses related to Design and Technology (ACARA, n.d.; School Curriculum and Standards Authority, 2014b). Regardless of the focus or the term used, Design-related courses are typically practical in nature (Stables, 2015).

The secondary school Design course is aimed to develop a broad range of knowledge, skills and understanding in Design students. In Western Australia, the Curriculum Council (now the School Curriculum and Standard Authority) of Western Australia described the rationale of the Design course as (2010a):

> … the strategic development, planning and production of artefacts of visual and tactile communication. It deals with the effective and efficient communication of ideas, values, beliefs, attitudes, messages and information to specific audiences. … The course equips students with the knowledge and skills to understand and interpret design, and to competently develop, plan and produce functionally effective artefacts for the world of today, and the future. (p. 3)

The terms that the Curriculum Council used to describe the Design course emphasise the students' active role, hence highlight the need for authentic assessment to measure student achievement. This course description portrays the practical nature of the Design course which consequently exemplifies the practical and contextual nature of the assessment. Based on this description, the assessment task should show the evidence of the students' ability to strategically develop, plan and produce artefacts of tactile communication, to show a good understanding and interpretation of design principles, and to manifest those principles on their design products.

It is important, therefore, that the assessment of the Design course authentically measure student achievement in the whole Design process, parallel to Kimbell's (2007) argument on the Design and Technology course: "Since design & technology is an activity that is premised on bringing about change in the made world, then common sense suggests that the best way of assessing learners' capability in design & technology is to put them into an activity and see how well they do it" (p. 47). As the Design course was aimed to develop the students' ability to answer to design-related challenges, the assessment should be on the evidence on how the students do so. Aligned to the Curriculum Council's description of the Design course (2010a), assessment in the Design course should assess the whole design process, from the planning stage to the development and presentation of the final design artefact. It should assess how the students derived influences from other courses, philosophies and products to create their original and innovative design. It should also assess how they employed their understanding, technical skills and different techniques in the task, and how they communicated their design ideas. This kind of assessment, by its very nature, is authentic assessment.

The practical nature of the Design course, the contextual nature of the skills that the course needs to help the students develop, and the applicability of these skills to the real-world situation require that the assessment tasks in this course be practical, contextual, and applicable. In the Design course, therefore, as well as in all areas in education, authentic assessment is essential.

Visual Arts course

Historically, visual art education curriculum in school has been intimately associated with the current social, political and economic situation (Australian Curriculum Assessment & Reporting Authority, 2011). Formal visual art education has dated back to the industrial age in 1870s, when schematic drawing training was introduced in schools to answer to the need of draftsmen for the raising number of factories in England. Since then the purpose of visual art education has developed gradually into a creative endeavour as history advanced into the depression era, technology era and into the information era. It has progressed into an instrument to promote visual appreciation, effective communication, creative thinking and innovation, and it has continued to be an inseparable and a valuable part of children’s education (Davis, 2008).

As the formal visual art education programs progressed, the need for assessment also grew. School accountability involves the evaluation of school programs and assessment of student progress. School has the responsibility to provide evidence of student academic progress, including in Visual Art. However, being a subject that celebrates creativity, uniqueness, inventiveness, imaginativeness and expressiveness, Visual Art and assessment are often viewed as contradictory (Eisner, 2002; Rayment & Britton, 2007; Taylor, 2006).

Traditionally, assessment is based on a standard. Art making, on the other hand, defies standards (Rayment, 2007). The primary concerns over the assessment in Visual Arts were over how to measure creativity and the fear that assessment would hinder creativity (Dorn et al., 2004). Correspondingly, Ewing (2010) emphasised the difference between arts being an emotional expression and assessment being a cognitive expression by explicating:

> Aesthetic knowledge is central to learning, understanding and enabling in our society. However, providing aesthetic knowledge is difficult for schools and teachers, because it is an experience that engages the brain, body and emotions, all together in a range of a symbolic languages and forms, whereas orthodox schooling and particularly assessment systems concentrate on those cognitive aspects of knowledge that can be made explicit and learned propositionally, just in words or numbers. (p. vi)

This perceived misalignment between art and assessment creates apprehension among art educators towards art assessment, especially when the assessment is high-stakes (Beattie, 1997; Rayment & Britton, 2007). Eisner (2002) pointed out several problems in using a standardised assessment tool such as a rubric in Visual Arts. The first problem was the practicality of an assessment standard. While the standard needs to be designed to be as detailed and diverse as possible, the more detailed and diverse it is, the more time and effort is required from the assessor. The second problem was the difficulty to describe arts in words. Rubrics need to be as descriptive as possible, however language is too limited to describe arts. The third was the different rate at which students learn. In arts, this is even more pronounced since fundamentally the arts are about individual expression and this encourages uniqueness.

Equity was also found to be another issue in art assessment. Findings from research conducted by Blaikie, Schönau, and Steers (2003) included different inclinations between female and male art students in terms of their likeliness to understand assessment criteria, expected achievement, and seeking feedback from their peers and teachers. To further complicate matters, certain societal structure such as patriarchy in the art world was also found to influence the way male and female students regard art education, achieve in art assessment, and consider a future in arts. Beside gender, students' socioeconomic background often determine their access to art resources and exposure (Ewing, 2010). Hence, students from low socio-economic background could be disadvantaged in art assessment. Equity issues such as gender and background could further muddy the water that is art assessment.

The awareness of an assessment could make the students align their art making to the assessment in order to attain a good result and sacrifice creativity for the sake of the assessment (Madeja et al., 2004). Furthermore, assessment requires a standard, and when there is a standard upon which assessment is based, there is also a concern that teachers would teach to the test (Rayment & Britton, 2007; Taylor, 2006). This would further skew the purpose of art education and limit creativity.

The main problem in art assessment is, therefore, in quantifying student progress or achievement, which raised the question of whether or not it should be measured (Dorn et al., 2004). As Eisner (2002) argued, however, these problems did not mean that art education should not involve assessment. On the contrary, these problems made it even more important for art educators to keep on trying to find the most suitable form of assessment based on the purpose of the assessment. Dorn et al.’s (2004) study regarding art teachers’ attitude towards assessment also indicated that the majority of art teachers considered assessment to be very important in arts education.

Common assessment practice in Visual Art education includes different types of tasks and artefacts such as portfolio, art journal, project-based learning and artwork exhibition (Beattie, 1997; Lockee & Wang, 2014). While there could be a wide variety of types of assessment task, for summative assessment in secondary school the most common ones are in the forms of art portfolio and finished artworks. Eisner (2002) identified three general features to find when assessing student artwork, which were the technical quality, the ingenuity of the idea and the aesthetic quality of the artwork. These three features are usually broken down into a more detailed description in an assessment rubric.

Similar to the Design course, the Western Australian Visual Arts course is practical and closely related to the real world, even though in Visual Arts the complexity is due to the creativity and aesthetic components more prominently than in Design. Also similar to Design, the nature of Visual Arts course makes it important that authentic assessment is used. Assessment tasks in both the Design and Visual Arts courses need to be practical and contextual for them to best represent student achievement. This need to assess creative expression through practical and contextual activity may lead itself to digital forms of assessment (Dillon & Brown, 2006; Doug, 2005; Jones-Woodham, 2009).

Task Assessment

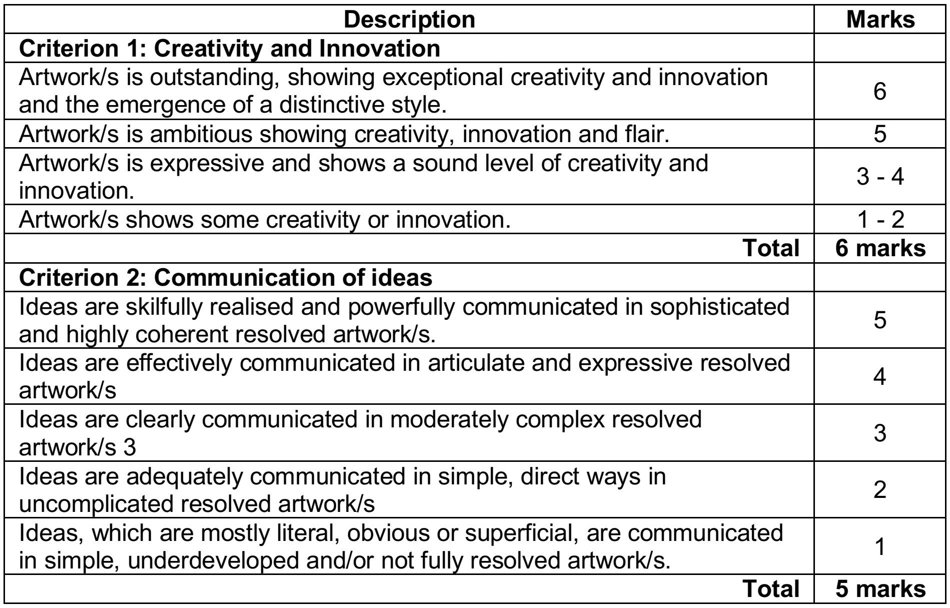
The Task Assessment component of the assessment processes pertains to matters related to the assessor side of the assessment process in measuring the performance or the skill and/or knowledge of the student in this study. In this component are processes such as scoring methods, which include the Analytical marking and Comparative Pairs judgements methods, training of assessors, and assessment quality, which include score reliability and assessment validity.

Scoring methods

Most task assessment in school today involves quantification that results in scores or grades. This quantification of educational achievement dates back to the late 1700s when the gradual rise of the need for the use of numbers driven by the mechanical inventions and foreign trade finally influenced assessment in education (Madaus & O'Dwyer, 1999). Numerical representation of correct answers subsequently started to be used to rank examinees, increasingly replacing examiners' qualitative judgements that were increasingly found to be subjective, and thus unreliable.

As society progressed and education became more and more accessible to the general public, the need for a more feasible and accurate measurement of education achievement also grew. Norm-referenced assessment was officially introduced in early 1900s in the Boston public school system (Madaus & O'Dwyer, 1999) with an aim to rank students based on their academic achievement. In norm-based assessment, students' scores were scaled to fit the normal distribution (Knight, 2001). While this approach might provide a result that was relatively unaffected by various external variables, the result was also lacking in information, for example the information on how the student achievement and test difficulty changed over the years (Sadler, 1987) because in norm-based assessment, the measurement model is the normal distribution and the result is always fit to the model.

The move from norm-based assessment towards criteria-based assessment started in the late 1970s, and with it the use of analytical marking method was becoming more and more common as the main characteristic of the *criterion-referenced* or *domain-referenced assessment* (Rust & Golombok, 2014; Sadler, 1987). In analytical marking method, the scoring process is conducted based on a set of criteria or standards that reflects the levels of achievement that are described in a marking rubric or even a simple marking key (Moskal, 2000). These criteria were created based on an analysis of achievement or the assessment task. An example of a rubric is as shown in *Figure 2.3*.



Description	Marks
Criterion 1: Creativity and Innovation	
Artwork/s is outstanding, showing exceptional creativity and innovation and the emergence of a distinctive style.	6
Artwork/s is ambitious showing creativity, innovation and flair.	5
Artwork/s is expressive and shows a sound level of creativity and innovation.	3 - 4
Artwork/s shows some creativity or innovation.	1 - 2
Total	**6 marks**
Criterion 2: Communication of ideas	
Ideas are skilfully realised and powerfully communicated in sophisticated and highly coherent resolved artwork/s.	5
Ideas are effectively communicated in articulate and expressive resolved artwork/s	4
Ideas are clearly communicated in moderately complex resolved artwork/s 3	3
Ideas are adequately communicated in simple, direct ways in uncomplicated resolved artwork/s	2
Ideas, which are mostly literal, obvious or superficial, are communicated in simple, underdeveloped and/or not fully resolved artwork/s.	1
Total	**5 marks**

Figure 2.3 A sample of rubric from the 2011 Visual Arts WACE examination marking key (Curriculum Council of Western Australia, 2010b, p. 3).

It is also common that the criteria that would be used as the base to score student work be made known to the students (Brookhart, 1999; Stowell & McDaniel, 1997) as an effort to increase the fairness of the assessment, especially in high-stakes assessment. With knowledge of the criteria, the students could focus their learning and achievement towards the direction intended in the course. The downside is that there is the possibility that, parallel to the teachers who would teach-to-the-test, the students would learn-to-the-test, and therefore miss the broader and more meaningful learning experience that they could have (Isaacs, 2013). One can argue, however, that in high-stakes assessment this particular threat is overshadowed by the importance that the students know how to best showcase their work or perform to achieve a high score (Jennings & Bearak, 2014).

Beside analytical rubrics, holistic criteria are also quite commonly used in education assessment, especially in performance assessment. Unlike analytical rubrics in which the final score is the total of scores from each level of achievement in each criterion, a holistic criterion is used to assess the overall quality of student work and assign a score or grade to represent that overall quality (Jonsson & Svingby, 2007; Moskal & Leydens, 2000; Perlman, 2003). An example of the use of a holistic criterion is allocating a grade A, B, C, D, or E and a comment to an essay.

The Comparative Pairs judgements method is a scoring method that dates back to the early 1900s but was recently made more feasible by developments in computer technology . In the Comparative Pairs judgements method usually the judgements are conducted based on a holistic criterion. In this kind of assessment method the assessors are given a pair of student’s work and they judge the superior one between the two, based on that holistic criterion.

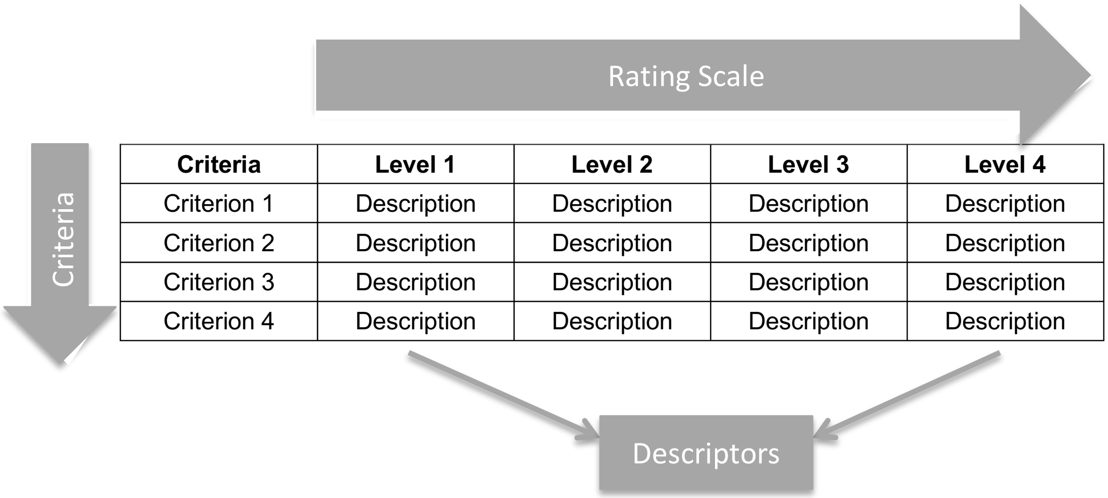
In the present study the Comparative Pairs judgements method was investigated by using the commonly used analytical marking method as a comparison. The analytical marking method employed used an analytical marking rubric. The Comparative Pairs judgements method employed an holistic criterion. In the next two sections each method is discussed within this limited context.

Analytical marking

The move towards authentic assessment approach in the late 1970s signalled the beginning of the use of criterion-referenced assessment as a preferred alternative to the conventional norm-referenced assessment (Madaus & O'Dwyer, 1999; Sadler, 1987). Because authentic assessment aims to assess student work that is more practical and meaningful than the traditional objective assessment (e.g., tests), it calls for a marking method that could assess quality (Jonsson & Svingby, 2007). The most common method that was considered appropriate was the analytical marking method. The analytical marking method is a method of marking student work based on pre-set standards or criteria. In analytical marking method, the assessment task is essentially analysed in terms of what is to be measured. The skills or knowledge aimed to be measured are represented in these standards or criteria (Andrade, 1997; Humphry & Heldsinger, 2009)

The analytical marking method is the most common marking method currently used in criterion-referenced assessment. Criterion-referenced assessment gained popularity after the previously used norm-referenced assessment was considered to be lacking in data on student achievement (Sadler, 1987). Because norm-referenced assessment is focussed on ranking student achievement on a normal distribution, it does not record information on the real data of student achievement, and consequently, on how the achievement fluctuates over the years. On the other hand, criterion-referenced assessment is more focussed on measuring student achievement based on a set of criteria or achievement standards. In the analytical marking method, these criteria or standards could be specified in various forms ranging from a simple marking key to a complex and detailed marking rubric. Student work is judged separately based on each criterion and the scores are then combined to an overall score or a grade (Sadler, 2009).

Marking rubrics usually contain three components, which are criteria, mastery level and descriptors (Andrade, 1997; Humphry & Heldsinger, 2009; Jonsson & Svingby, 2007). Achievement criteria explicate specific knowledge domains within the learning objectives, for example originality of artwork or presentation of design project. Mastery level breaks down the progression of student achievement into several levels, for example fail, pass, meets expectation and exceeds expectation. Descriptors describe the quality of work that constitutes each level in each criterion. *Figure 2.4* shows a matrix rubric adapted from Gallaudet University (n.d.) based on these components.



Criteria	Level 1	Level 2	Level 3	Level 4
Criterion 1	Description	Description	Description	Description
Criterion 2	Description	Description	Description	Description
Criterion 3	Description	Description	Description	Description
Criterion 4	Description	Description	Description	Description

Figure 2.4 Components of a marking rubric (adapted from Gallaudet University, n.d.).

Popham (1997) delineated the role of rubrics as “not only scoring tools but also, more important, instructional illuminators” (p.75). Rubrics could help teachers planning their lessons and evaluating student learning, as well as help students planning and evaluating their own learning (Andrade, 2005; Jonsson & Svingby, 2007; Popham, 1997). However, as Popham further explicated, even though the design quality of rubrics influences assessment quality, unfortunately rubrics are not always well designed. Early on Popham suggested that rubrics should be brief, task-appropriate and practically teachable (1997). Later, Jonsson and Svingsby (2007) suggested that reliable rubrics are “analytic, topic-specific, and complemented with exemplars and/or rater training” (p. 130) and recently Humphry and Heldsinger (2009; 2014) warned against halo effect created by semantic overlaps of criteria in rubrics.

The use of a rubric in assessment is often considered constructive. Rubrics provide both teachers and students with explicit feedback to assist with planning, teaching, learning and evaluating (Popham, 1997; Stowell & McDaniel, 1997). In performance assessment the use of good quality rubrics could increase the reliability of the scores and the validity of the judgement (Jonsson & Svingby, 2007). In high-stakes assessment in particular, this increases the accountability factor of the assessment. Several statistical models such as the Rasch polytomous model and Item Response Theory (IRT) are currently broadly used to analyse the quality of test items, the quality of rubric design, and student attainment

(Andrich, 1988; Bond & Fox, 2001; De Ayala, 2009; Fischer & Molenaar, 1995; Ostini & Nering, 2006). For example Humphry and Heldsinger (2009) used the Rasch polytomous model to test the quality of narrative writing rubrics.

Comparative Pairs judgements (with holistic criterion)

Human judgement is relative, limited, and susceptible to change. It is restricted by natural cognitive limitation and it depends on background and experience. Laming (2011) made a premise on human judgement, “There is no absolute judgment. All judgments are comparisons of one thing with another” (p. 9). Laming further explicated that relatively accurate judgements are only possible when they are made in rank order, regardless whether it is judgement on physical properties such as temperature and speed or judgement on psychological traits such as intelligence and attitude. Thurstone (1928) called this judgement process “discriminal process” (p. 274). It is essentially a process of discriminating the quality of an object or a concept in terms of which one is better than the other, and this process was considered to produce a more reliable result than assigning a score to represent the quality of what is being measured.

Judgements on several physical properties can be made easier and more accurate with measuring tools such as a ruler to measure length and a scale to measure weight, but the measurement process is still a comparison of the property with a standard, which is relatively constant. In judgements of psychological traits, however, there cannot be a constant standard because psychological traits are obscure (Angoff, 1996; Laming, 2011; Thurstone, 1928).

As in judgements of psychological traits, the judgement of authentic assessment tasks such as the creation of an artwork in Visual Arts and the development of a portfolio in Design, tends to be highly subjective. Therefore, the construction of a standard and the judgements made against the standard could be problematic. This standard is usually based on the previous student achievement data collected (Angoff, 1996; Karantonis & Sireci, 2006). Consequently, equity could be an issue in the use of achievement standards when the student demographics have changed (Andrade, 2005).

Unlike the commonly used analytical marking method (e.g., rubrics), in Comparative Pairs judgements method assessors do not assign scores to student work. Instead, they simply compare two works and make a judgement on which one is better based on a holistic criterion. Thurstone first introduced this judgement method in 1927 in his paper *A Law of Comparative Judgment*. In his paper, he argued that Weber's Law and Fechner's Law on human perception to discriminate the difference between two physical stimuli could also apply in psychological measurement. Similar to how we could make judgements on the disparities on physical properties among physical objects, he asserted that this discrimination process could create a measurement scale for psychological traits, academic attainment included.

With the advancement in computing technologies, the application of Thurstone's theory has become easier (Bartholomew & Connolly, 2017). A variety of computer software, such as the Adaptive Comparative Judgement (ACJ) (Pollitt, 2012b) and No More Marking ("*No More Marking"*, 2017) systems which are based in the United Kingdom, and Brightpath ("*Brightpath*", 2017) that is based in Perth, Western Australia; have been created to manage both the student digital work and the judgement process. The calculation of scores requires iterative algorithms that analyse the wins and losses data. Current technologies allow for efficient and economical implementation of this task assessment method including the development of the pairing, the scaling of the s and the analysis of the quality of the judgements. A further discussion on the ACJ system used in the present study is presented in Chapter 3.

Many studies on the Comparative Pairs judgements method in education have been done in different contexts in different countries, such as the United Kingdom, Singapore, United States of America, and Australia. In the Comparative Pairs judgements method, student works are paired for judgements. Judgements are usually based on a holistic criterion of quality required for the assessment, which in some cases could be quite general; for example conceptual calculus understanding (Jones & Alcock, 2012), narrative writing quality (Heldsinger & Humphry, 2010), Geography essay quality (Pollitt & Whitehouse, 2012) and chemistry investigation quality (McMahon & Jones, 2015); or specific such as

"Prototype product is effective for target customers through developed planning to incorporate all the required features and information, appropriate use of aesthetic effects on a theme, consistent and balanced layout, and professional look" (Newhouse et al., 2011, p. 63) for Applied Information Technology (AIT) digital portfolio.

In performance assessment in particular, the Comparative Pairs judgements method with a holistic criterion is considered valuable (Jones & Alcock, 2014; Kimbell et al., 2009; Pollitt, 2012b), mostly because performance assessments are subjective and complex. In subjective and complex assessments, assigning a score to a quality such as is done in analytical marking is potentially problematic because of the factors of assessor bias and assessor leniency, which would affect the reliability of the scores. Gill and Bramley (2008) argued that it was likely that relative judgements are more accurate than absolute judgements. With complex assessment tasks requiring subjective task assessment the possibility that absolute judgements are less accurate is even more likely.

Data analysis in the Comparative Pairs judgements was based on Thurstone's law of comparative judgement equation that was adapted into a Rasch dichotomous model (Andrich, 1978; Whitehouse & Pollitt, 2012). The following section discusses this logistic model.

The Rasch model for Comparative Pairs judgements

The advantage of objective measurement such as IRT and the Rasch models over raw scores is as Tatum (2000) summarised "Observational statistics like raw scores and ratings describe a one-time event with all elements interwoven" (p. 274). Both methods were based on the premise that elements within educational measurement, or facets, do not have absolute values. Raw scores are not absolute. They are a product of elements such as difficulty of items, assessors' judgements, and scaling process. Variations from these elements and the interactions among them make raw scores arbitrary. On the contrary, as Tatum further attested "Objective measurement gives us straight lines, precise measures, and separated elements that remain stable across time and samples" (p. 274).

The Rasch Model was developed by Georg Rasch in the 1960s (Andrich, 2004; Bond & Fox, 2001). It has since been widely used in many areas such as health, psychological measurement and educational measurement. Unlike observational statistics, the Rasch model provides objective measurement that takes into account the ordering of item difficulty and person ability, the relationship between those two, and misfit data, all in one continuum, or scale, with equal units (Bond & Fox, 2001; Cavanagh & Waugh, 2011).

Item Response Theory (IRT) and the Rasch model were two model-based measurements that replaced the Classical Test Theory (CTT) in mid 1960s. CTT simply formulated the observed score as a linear function of true score plus a random error; it did not include variables such as test item difficulty and variations of score distribution (De Ayala, 2009; Fischer & Molenaar, 1995; Yu, 2011). CTT is formulated as:

$$X_i = T_i + E_i$$

where X is the observed score, T is the true score, and E is the random error.

On the other hand, IRT and the Rasch model calculate these variables and provide a probabilistic analysis of both item difficulty and person ability parameters. The Rasch simple logistic model, or the Rasch dichotomous model, (Andrich, 1978; De Ayala, 2009; Kimbell, 2008; Pollitt, 2012b) formulated the probability of a person being successful such as giving a correct answer (coded as 1) on an item as:

$$p((x_j = 1 | \theta, \delta_j) = \frac{e^{(\theta - \delta_j)}}{1 + e^{(\theta - \delta_j)}}$$

where θ is the person location parameter and δ is the item parameter for item *j*. The values for the probability are ranging from 0 to 1, with a value of 0.5 obtained when the value of θ is equal to δ, which is the condition when the person of θ ability has a 50% chance of being successful for an item of δ difficulty. Both the variables θ and δ are ranging between -∞ and ∞, from low ability and difficulty to high ability and difficulty. This probabilistic model is similar to a special case of IRT and is sometimes referred as such even though both models were developed separately (Andrich, 2004; Embretson & Reise, 2013).

In the analysis of judgements obtained from the Comparative Pairs scoring, the Rasch dichotomous model is simplified into a logistic function of odds (Whitehouse & Pollitt, 2012, p. 2):

$$\log odds\ (A\ beats\ B|v_a - v_b) = \ v_a - v_b$$

where v_a and v_b are the perceived quality of objects A and B, which refers to the criterion used in the judgements. This equation states that “the difference between the perceived quality of A and the perceived quality of B is equal to the log of the odds that object A will be judged to be better than object B”. As each pairing is judged in the Comparative Pairs judgements process, the difference between the perceived quality of each object, or student work, with another object creates an estimated score. This estimated score is in logits.

The reliability coefficient of these scores is obtained through a conversion of the separation coefficient into a coefficient that is similar to Cronbach’s alpha coefficient (Newhouse et al., 2011, p. 63; Whitehouse & Pollitt, 2012):

$$\propto = \frac{G^2}{(1+G)^2}$$

G is the separation coefficient, which is the ratio of the standard deviation of the parameter values and the root mean square of the estimation errors.

$$G = \frac{sd_v}{rmse}$$

Hence, the reliability coefficient is estimated from the spread of the location parameters of the portfolios as well as the consistency of judgements from the assessors. Therefore, this coefficient represents both the internal reliability of the scores and the inter-rater reliability among assessors.

Assessment quality

Assessment quality refers to the expected characteristics of the outcomes of the assessment process. Such characteristics are reliability, validity, authenticity, and accountability (Campbell, 2008; Elliott, Compton, & Roach, 2007; Miller, 2011). The quality of assessment depends on both the quality of the assessment task and the quality of the task assessment (Campbell, 2008; Miller, 2011). In this study, only reliability and validity were used to define the quality of the assessment, because of their relevance and also because of the limitations of the context and scope of this study. This section discusses these attributes and how these attributes are pertinent to data analysis and findings.

Reliability

Anastasi and Urbina (1997) defined reliability as “the consistency of scores obtained by the same persons when they are re-examined with the same test on different occasions, or with different sets of equivalent items, or under other variable examining conditions” (p. 84). Simply put, reliability is a measure of consistency or reproducibility; or as Thorndike and Thorndike-Christ (2010) expounded, reliability is about asking the question, “How accurately will the score be reproduced if we measure the individual again?” (p. 119). Reliable test scores should place students in relatively similar positions within the same group in a different set of test scores.

The main purpose of the measure of reliability, as Anastasi and Urbina (1997) explicated further, is to indicate how much the differences between scores of the same student were caused by chance and not by other reasons that might cause concerns. When there is a notable difference between the positions of scores obtained by a student in the first task and the second task, which would result in a relatively low reliability coefficient, the assessment developer needs to examine the possible reasons that could cause that difference. When the reliability coefficient was sufficient, then the variability might likely be due to chance. Several factors that could affect the reliability of assessment task result are Thorndike (1997):

- quality: length, difficulty, discrimination, wording
- conditions: time limits, instructions, physical environment
- cohort characteristics: range and distribution of ability.

While it is not possible to have an ideal assessment, developers and administrators have the responsibility to ensure that the effects of these factors are at minimum. Therefore, reliability measure needs to be estimated as the first indication of the level of consistency of result. There are several statistical methods to estimate reliability, depending on the types of assessment task and error to be measured. One of the most common methods is coefficient alpha, or Cronbach's alpha, which can be used to calculate the reliability coefficient of many types of test (Frisbie, 1988; Traub & Rowley, 1991) by comparing the correct answer of each student with the statistical spread of the result of the cohort. Several other methods can only be used for certain types of test or purpose, for example K-R20 and K-R21. Regardless of the method, the reliability coefficient has a range between 0 and 1, with 0 representing complete error and 1 representing the ideal condition of perfect reliability. The recommended reliability coefficient is around 0.50 for teacher-made tests and around 0.90 for standardised tests (Frisbie, 1988; Miller et al., 2009). There are several methods to estimate the reliability of a set of test scores, depending on the type of reliability to be measured. These methods are described in Table 2.2.

Table 2.2

Methods of Estimating Reliability (Frisbie, 1988; Kubiszyn & Borich, 1993)

Method	Measure of	Procedure
Test-retest	Stability / consistency over time	One test Same group Different time
Equivalent forms	Equivalence / consistency between two tests	Two tests Same group Close succession
Test-retest with equivalent forms	Stability and Equivalence	Two tests Same group Different time
Split-half	Internal consistency	One test Two equivalent halves of cohort
Coefficient alpha	Internal consistency	One test No split-half
Inter-rater	Consistency between assessors	Same test Same group Two or more assessors

The first five methods of reliability test are about detecting errors that are related to the assessment task, such as a test. For example, the test-retest method is usually suitable to test the consistency of the test result as caused by the consistency of the quality of the instrument over time. While more assessors could potentially increase the validity of test result, having more than one assessor to assess student work also add a threat to the score reliability. Aside from the consistency of the instrument, with more assessors there arises the issue of consistency between or among assessors. The inter-rater reliability measures this consistency.

Several research studies suggested the use of a detailed scoring rubrics to increase inter-rater reliability (Miller et al., 2009). Baird, Greatorex, and Bell (2004); (Brookhart & Chen, 2014; Jonsson & Svingby, 2007; Stemler, 2004) identified several factors that could influence inter-rater reliability. Beside the use of a detailed scoring rubric, these factors included assessors' quality and collaboration, the use of a standard and the quality of the test items. Professional learning aimed to "develop shared interpretations of assessment tasks and the requisite standards, especially through moderation, and to develop a common language for describing and assessing students' work" (Wyatt ‑ Smith, Klenowski, & Gunn, 2010, p. 72) was also considered to increase inter-rater reliability. According to Stemler (2004), there are three types of inter-rater reliability measurements, which are consensus estimates, consistency estimates and measurement estimates as summarised in Table 2.3.

Table 2.3
Different Types of Inter-rater Reliability (Stemler, 2004)

Inter-rater Reliability Estimate	Estimation of	Statistic Methods
Consensus estimate	The exact agreement between assessors in their interpretation of the scoring rubric	Percent agreement Cohen's kappa
Consistency estimate	The consistency of how the assessors assign scores based on the scoring rubric	Pearson's r Spearman's rho Cronbach's alpha
Measurement estimate	The accumulation of information from all judgements to develop a model to estimate the final score for each student	Principal component analysis Generalizability Theory Facet rater severity indices and fit statistics

These three types of inter-rater reliability estimate measures are different in terms of the severity of the agreement, the focus of the measurement of the agreement between assessors and the complexity of the statistical methods and information resulting from the computation. Consequently, the measurement of inter-rater reliability should take into account the type of data being measured and the purpose of the measurement.

In the present study, the whole Comparative Pairs judgements process was conducted within the ACJ system. At the end of every judgement round, the ACJ system uses a Rasch dichotomous model to calculate the reliability of the result of judgements. This system is discussed in Chapter 3, including the way the system calculates the reliability coefficient of the judgement results. The Rasch model that is employed by the ACJ system measures reliability for both the students (persons) and the items. However in this study, only the reliability estimates of the students, or person reliability index (Stemler, 2004), which concerns the consistency about the ordering of students based on their ability, is discussed because the Comparative Pairs judgements method used a holistic criterion. At the same time, the consistency estimate of the inter-rater reliability for the Analytical marking was calculated using a Pearson correlation coefficient, as is discussed in the Methodology chapter.

The errors of measurement are indicated by the inconsistency reflected in reliability testing. Anastasi and Urbina expounded that reliability tests distinguish these errors as "'true' differences in the characteristics under consideration and the extent to which they are attributable to chance errors" (1997, p. 84). Parallel to this, Frisbie (1988) defined reliability as "the property of test scores that describes how consistent of error-free the measurements are" (p. 25) and categorised these errors as systematic and random errors. While all measurements would comprise a certain level of random errors, reliability coefficient does not reflect systematic errors. Random errors reflect variations in test scores obtained by the same individual from a reliability test that are caused by chance. In comparison, systematic errors usually hold a pattern, and therefore, predictable. Systematic errors could influence either the whole cohort of students or individual students. Because both types of error could affect the interpretation of student score, both affect the validity of the measurement. Consequently, it is important that even with a set of scores with high reliability coefficient, an analysis of possible systematic error is also conducted. In the present study an analysis of discrepancies between methods of scoring was conducted.

Validity

In essence, validity is concerned with how well an assessment measures the constructs it sets out to measure (Burton, 2006; Frisbie, 1988). Consequently, every aspect of an assessment closely influences validity. The entire process of an assessment; starting with the planning stage and followed by the development stage, the implementation stage, the scoring stage, the interpretation stage and the reporting stage; affects validity. The main focus of assessment validation, however, lies on the interpretations and consequences of the assessment (Gipps, 1994).

Types of Validity

The history of assessment validity dates back to test validation in the 1920s when criterion validity was the standard of test validation (Kane, 2006; Messick, 1987, 1990, 1993a, 1995; Sireci, 2007). This type of validity is concerned about how well a test correlates to an external standard, or a set of criteria (Kane, 2006; Messick, 1993b). The use of criterion validity to validate a test could be difficult in some circumstances, for example when there is a lack of a suitable set of criterion to be used as a comparison. Today there are many types of assessment validity that are considered, including content validity, construct validity, concurrent validity, and consequential validity (Cizek, Koons, & Rosenberg, 2011; Moss et al., 2006), with content validity, criterion validity, and construct validity being the main types (Raykov & Marcoulides, 2011).

Another type of validation examines the content validity of a test. Content validity represents the degree to which a test measures the elements of a construct (Cureton, 1951; Kane, 2006). While criterion validity measures the relationship between a test and a set of external criteria, and therefore a quantitative measure, content validity is a qualitative, internal measure of a test. This validation method, however, serves as a prerequisite to criterion validity. If there is doubt about the content validity of a test, then there is little value to its criterion validity. The limitation of content validity is the subjectivity, and with that, the potential for conformity bias (Cureton, 1951).

Construct validity was initially introduced in 1950 as a validation method when other validation models were not possible (Cronbach & Meehl, 1955; Kane, 2001). Raykov and Marcoulides (2011) defined construct validity as “the extent to which there is evidence consistent with the assumption of a construct of concern being manifested in subjects' observed performance on the instrument” (p. 190). Construct validation is particularly necessary for assessments that measure latent traits, or traits that could not be directly observed, such as attitude, intelligence, or ability. Construct validity was subsequently accepted as the model that integrated validity theory based on the argument that construct validity encompassed the other two main validation models and was consequently applicable in most test situations, especially those in which the other two main models did not apply. This theory is known as the unified theory of validity (Kane, 2016).

The unified theory of validity was prompted by Loevinger in 1957 (Kane, 2001) and followed through by Messick (1989) when he proposed the definition of validity more comprehensively as,

> … an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores. … Broadly speaking, then, validity is an inductive summary of both the existing evidence for and the potential consequences of test interpretation and use. Hence, what is to be validated is not the test as such but the inferences derived from test scores – inferences about score meaning or interpretation and about the implications for action that the interpretation entails. (p. 5)

This definition is aligned with, and in some cases is the foundation of, the characteristics of validity outlined in literature. Messick (1987, p. 1) and other scholars on test validation delineated that validity:

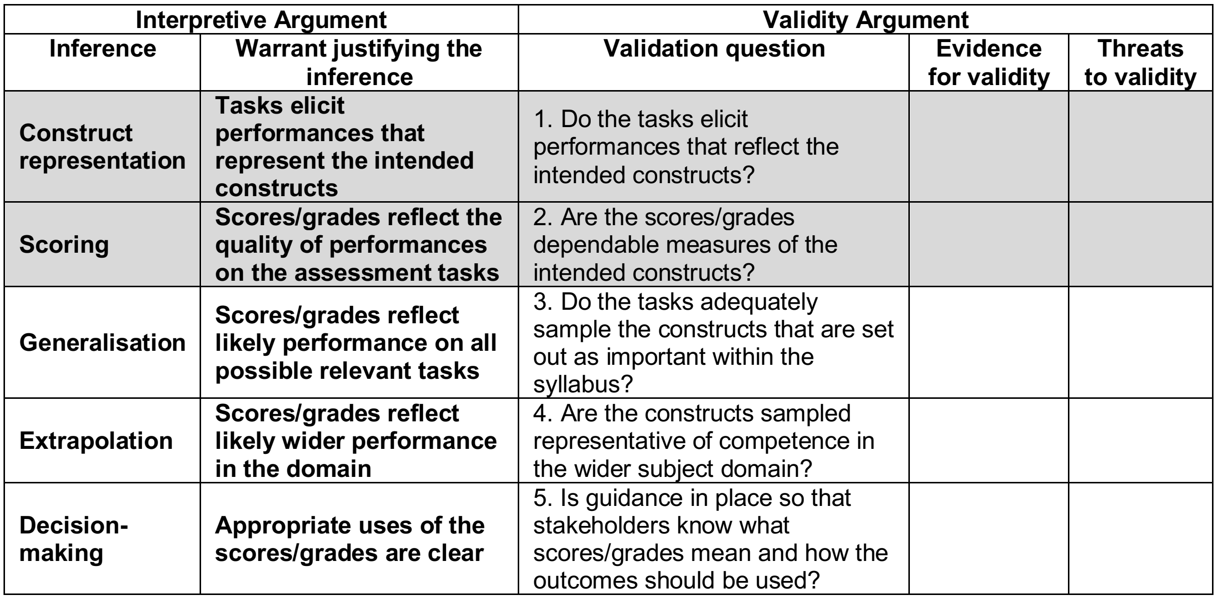
- is based on theory
- is built from evidence (Kane, 2004; Lane, 2014; Shaw, Crisp, & Johnson, 2012)
- is a continuous process, as new evidence for or against it is found (Cizek et al., 2011; Cronbach, 1971; Elliott et al., 2007; Kane, 2006; Raykov & Marcoulides, 2011; Shaw et al., 2012; Sireci, 2007)
- is a continuum, as opposed to a dichotomous concept of valid and not valid (Raykov & Marcoulides, 2011; Sireci, 2007)
- is inferred, as opposed to measured (Pollitt, 2012c)
- is a property of test scores instead of the test itself (Messick, 1996)
- is concerned with how appropriate the inferences based on the test scores are (Brookhart & Nitko, 2008; Messick, 1980; Raykov & Marcoulides, 2011; Sireci, 2009)
- can decline over time and process (Cronbach, 1971; Messick, 1980; Shaw et al., 2012; Thorndike, 1997).

Beside these characteristics, reliability is also a feature of validity (Cronbach, 1971; Kane, 2004; Pollitt, 2012c). Validity is concerned about the interpretations that could be made based on the assessment result, while reliability is a measure of the accuracy of the result. Therefore reliability is a major characteristic of validity, and was central to some of the analysis for the present study.

In psychological and educational assessment in particular, construct validity, and thus the unified theory of validity, was viewed to have an advantage because many assessments in these areas measure latent traits. The weaknesses of criterion validity and content validity meant there were circumstances in which they were not applicable or sufficient. Since the unified theory of validity relied upon various sources of evidence, it could be applicable in most types of assessment.

Evidence-Based Validity Analysis

Evidence-based validity analysis on an assessment could be conducted in two ways. The first approach is by building evidence for validity (Raykov & Marcoulides, 2011; Shaw et al., 2012), the second is by eliminating threats against validity (Cizek et al., 2011; Cronbach, 1971; Elliott et al., 2007; Kane, 2006; Messick, 1989; Raykov & Marcoulides, 2011; Shaw et al., 2012; Sireci, 2007). Shaw et al. (2012) proposed a validation framework that combined the two approaches based on Kane's work (Shaw et al., p. 167). The framework is as shown in *Figure 2.5*:



Interpretive Argument		**Validity Argument**		
Inference	**Warrant justifying the inference**	**Validation question**	**Evidence for validity**	**Threats to validity**
Construct representation	**Tasks elicit performances that represent the intended constructs**	1. Do the tasks elicit performances that reflect the intended constructs?		
Scoring	**Scores/grades reflect the quality of performances on the assessment tasks**	2. Are the scores/grades dependable measures of the intended constructs?		
Generalisation	**Scores/grades reflect likely performance on all possible relevant tasks**	3. Do the tasks adequately sample the constructs that are set out as important within the syllabus?		
Extrapolation	**Scores/grades reflect likely wider performance in the domain**	4. Are the constructs sampled representative of competence in the wider subject domain?		
Decision-making	**Appropriate uses of the scores/grades are clear**	5. Is guidance in place so that stakeholders know what scores/grades mean and how the outcomes should be used?		

Figure 2.5 Validation framework for written examination (Shaw et al., 2012)

Kane (2006) proposed a two-step validation framework that consisted of interpretive arguments and validity arguments. Interpretive arguments specified the intended inferences derived from test results while validity arguments assessed the appropriateness of those inferences. Shaw et al. (p. 23) developed Kane's framework to include *construct representation*, which gathered information on the validity of test items. In Kane's original version, this argument was included in *extrapolation*, along with the inferences about future studies. In the present study, Shaw's *construct representation* inference was preferred, to better define the distinction between the inferences of *construct representation* and *extrapolation*. Even though in this framework the term *test* was used, the present study took the assumption that this validation framework also applied to the broader concept of most types of assessment tasks including digital portfolio and artwork. This assumption was based on the non-specific terminology of the types of assessment task used in related literature regarding this framework (Kane, 2006; Shaw et al., 2012).

According to Shaw et al. (2012) validation on *construct representation* concerns itself with the appropriateness of the task with the performances it was designed to measure. Interpretive argument on *scoring* evaluated the appropriateness of the scoring process and criteria with the constructs being measured. *Generalisation* assessed the validity of the test results with reference to a broader construct domain of standard performance or achievement, or in another word, the extent to which the test results represented the performance or achievement as identified in the syllabus. *Extrapolation* questioned how well the test results correlate to the competency standard set for the course and further study or employment, and also outlined external factors that might contribute to score variability. *Decision-making* evaluated the appropriateness of the use of test results.

Kane (2006, p. 23) suggested several possible methods to gather sources of evidence of validity, or threats to validity. In the present study, only two interpretive arguments were considered relevant: *construct representation* and *scoring*, as is specified in the shaded sections on *Figure 2.5*.

Table 2.4 presents the validity arguments that were suggested by Shaw et al. (2012) for these arguments.

Table 2.4
Validity Arguments (Shaw et al., 2012, p. 167)

Validation question	Possible methods
Do the tasks elicit performances that reflect the intended constructs?	Review examiner reports for insights into how the questions were answered by candidates. Analyse performance data (e.g. item level scores for a sample of candidates using statistical methods (e.g. Rasch, factor analysis) to explore item functioning, relationships between items, and to check for test bias (e.g. using differential item functioning analyses by gender, school type, etc.). For misfitting items, analyse the nature of candidate responses to gather insights into any possible sources of construct irrelevant variance. Ask appropriate examiners/experts to rate the extent to which each question appears to elicit each assessment objective set out in the syllabus (using this as a proxy for the constructs). Ask appropriate examiners/experts to rate the extent to which each question places certain types of cognitive demands on students.
Are the scores/grades dependable measures of the intended constructs?	Review exam board documents on marking and scoring procedures. Ask a number of markers to mark the same exam scripts in a multiple re-marking exercise so that the consistency and reliability of marking can be analysed. Conduct statistical analyses of candidate exam results to explore issues relating to aggregation of test scores and intended and achieved weightings of exam components. Conduct composite reliability analysis. Statistical analysis of the effectiveness and accuracy of classifying students to grade bands based on marks.

The validity arguments that were used in the present study were adapted to suit the limitations and purpose of this study, as well as Samuel Messick's (1994) argument on the validation of performance assessment. In performance assessment, Messick differentiated the validation process of assessment based on the focus of the assessment. When the assessment task, whether it is a product or a process, is the target of the assessment, "All that counts is the quality of the performance or product submitted for evaluation, and the validation focus is on the judgment of quality" (p. 14).

Messick (1994) further stated that in performance assessment validity of the assessment can be defined in terms of the collective knowledge of the assessors. Consequently, assessor inconsistency could be the main threat to the validity of the results from the Comparative Pairs judgements process. Correspondingly, Pollitt (cited in Kimbell et al., 2009) proposed three possible sources of assessor inconsistency, which were: "(a) variation in absolute standard; (b) differences in discrimination or spread of marks; (c) differences in conceptualisation of the trait being measured" (p. 79). The first two sources of inconsistency came from the subjectivity in assigning scores that stemmed from differences in each assessor's standard. As an example, for a criterion that has a 1 to 6 score range, different assessors might have different opinions on whether an artwork could be associated with a 1 or a 2, or even to a larger score, because of differences in each assessor's standard. A different piece of artwork could be judged as 1 score higher, 2 scores higher, or even a larger difference, because of differences in how each assessor discriminated the scores. The third threat referred to the possibility that different assessors might translate evidence for certain traits differently, for example the level to which a Design portfolio shows a skilled design process. In Comparative Pairs judgements method this threat could manifest in a difference of which student's work is adjudged to be the winner.

The first two threats exist in analytical marking. A way to minimise these threats was the use of a well-designed analytical scoring rubric, and training the assessors to use this rubric to achieve a similar perception on how to interpret the rubric (Baird, 2007; Brown, Bull, & Pendlebury, 1997; Pollitt, 2012b; Reddy & Andrade, 2010). Scoring rubrics are especially useful for scoring subjective tasks such as performance tasks, portfolios, essays and artworks because they provide a descriptive, detailed guideline of the marking scheme (Brookhart & Nitko, 2008; Jonsson & Svingby, 2007; Moskal & Leydens, 2000). Furthermore, they also specify the constructs that should be assessed and the scores that should be assigned to a range of achievement levels within each construct, based on the purpose of the assessment. Training the assessors to use the rubric could further reduce the discrepancies that might be caused by assessors' differences in judgement.

The third source of inconsistency was the variations of assessors' interpretation of the criteria, which potentially affect the internal reliability of the scores, the inter-rater reliability, and the validity of the assessment (Kimbell et al., 2009). This source of inconsistency was applicable in both analytical marking and the Comparative Pairs judgements method. A way to achieve agreement in the way the assessors interpret the criteria was assessor training (Baird et al., 2004; Brown et al., 1997; Rayment, 2007).

Conceptual Framework

The conceptual framework that guided this study was based on the adapted version of Campbell's (2008, p. 57) and Miller's (2011, p. 34) frameworks on assessment process. *Figure 2.6* depicts the elements of the assessment process and the relationship between those elements. The shaded areas highlight the elements investigated and discussed in this study, in contrast with the broader framework investigated in the project. In the discussion that follows, the terms used in the diagram are in italics.

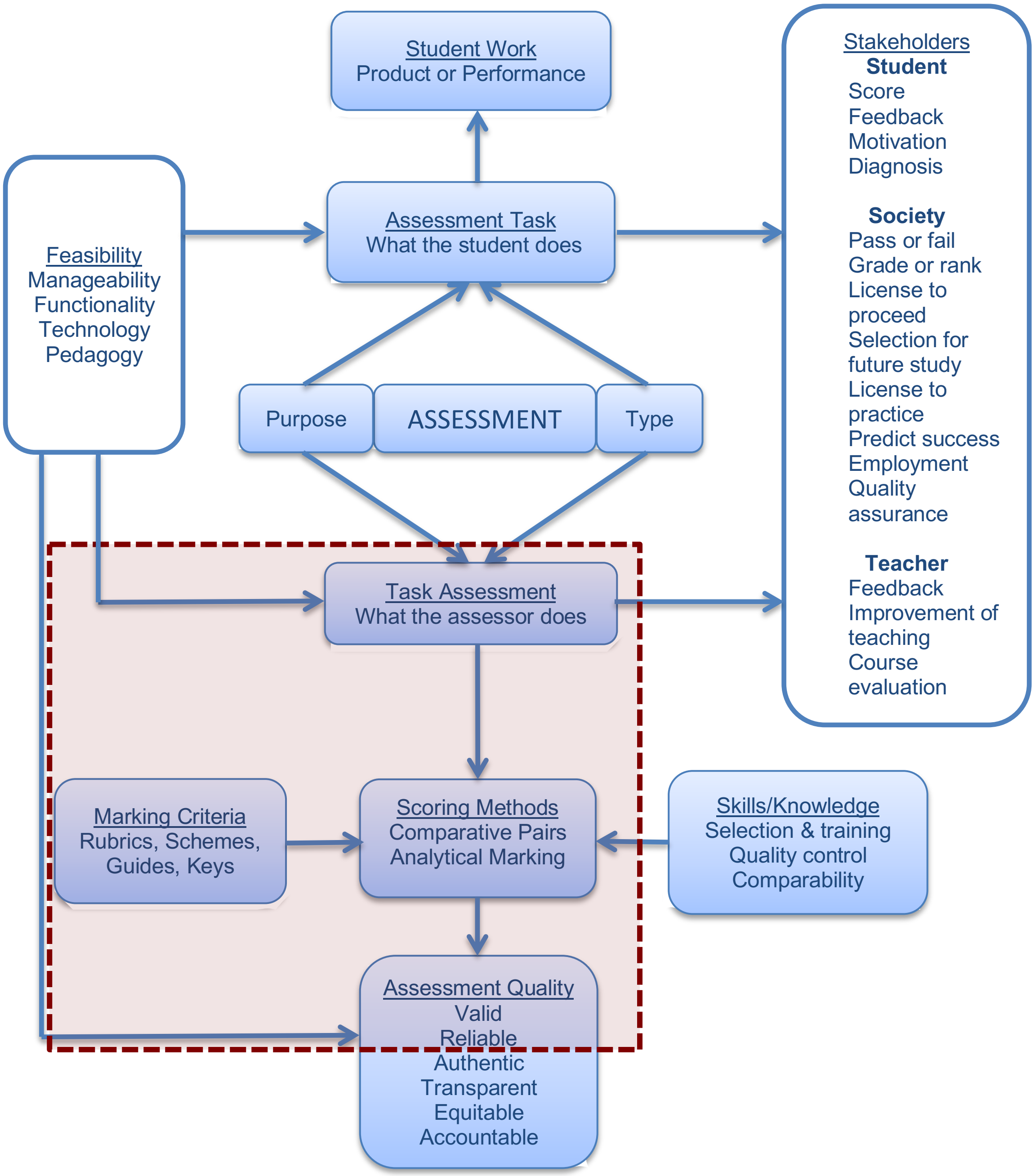


Figure 2.6 The assessment process framework.

This concept of the assessment process consists of a broad range of elements. This study was focussed on a select few of the elements that were directly connected to the utilisation of the Comparative Pairs marking. The shaded box represents the elements within the assessment process framework that were investigated and discussed in the present study. *Figure 2.7* shows the conceptual framework of the present study, which was built from these elements.

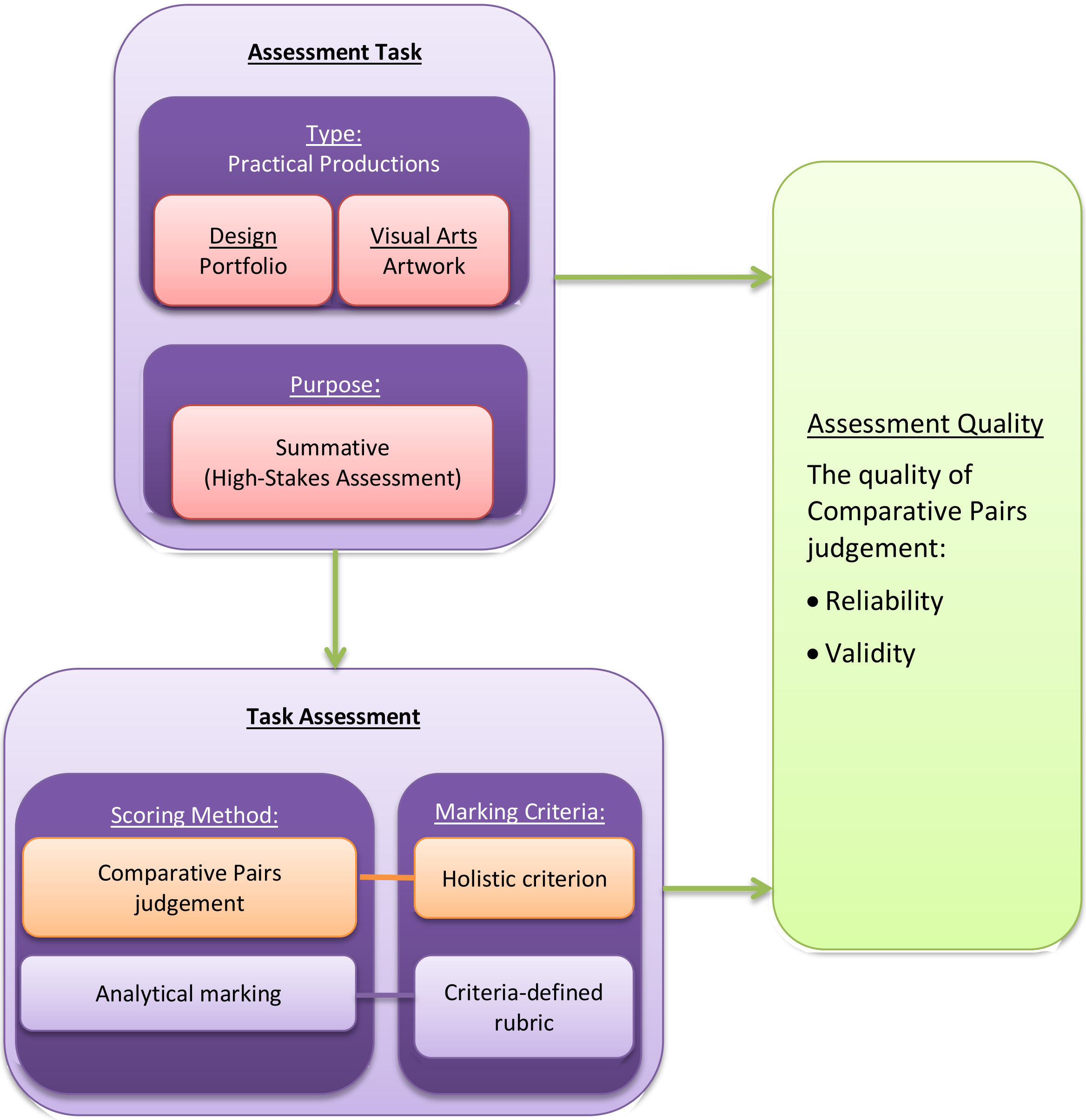


Figure 2.7 The Conceptual framework.

The conceptual framework of the present study observes the assessment process as consisting of two processes: the *assessment task* and the *task assessment*. The *assessment task* is associated with the task factors such as the type of *student work* required for the assessment and how *feasible* the execution of that type of work is. The *task assessment* is associated with the assessor side of the assessment process, which includes the *marking criteria* used in the marking process, the *marking methods* employed, and the *skills and knowledge of the assessors*. In this study the *assessment task* part of the framework only provided the background for the *task assessment* part, therefore only *assessment task* elements pertained to this study was discussed. Elements such as processes and theories related to test development were not discussed in depth. The construct validity of the test was assumed, considering WACE is a high-stakes, standardised test. The focus of this study was more on the *task assessment* with a discussion on how different *assessment tasks* might have affected the quality of the *task assessment*.

Two interrelated factors of assessment, the *purpose* and *type* of assessment, basically determine the design of an assessment, and, both the assessment task and the task assessment. In this study the assessment investigated was the practical component of a summative assessment, which was the WACE examination. While the type of assessment task for both subjects were practical productions, the specific of these tasks were different. The *assessment task* was designed to suit this purpose. The *student work* submitted for the WACE examination for the two secondary school subjects investigated was in the form of a finished product. In the Design course a portfolio consisted of the evidence of the development of a Design project was the requirement. Quite differently, an artwork accompanied with an artist statement was the *student work* required in the Visual Arts course.

The *purpose* of this examination was to assess student achievement at the end of a program. Both *types* of assessment, Criterion Referenced Test (CRT) and Norm Referenced Test (NRT) were used to analyse the student level of achievement. Results from the Comparative Pairs judgement is based on a holistic criterion (in this study), but because it places each student in a location parameter that is relative to the cohort, it is also norm-referenced (Bond, 1996; Brown et al., 1997; Burton, 2006). However, unlike in a norm-referenced test, the location parameters do not follow a pre-determined cut-off and even though they resulted from comparisons between student works, they are not concerned with a certain distribution pattern but their own. This makes this scoring method unique.

Similar to *assessment task*, the *task assessment* also depends on the *purpose* and *type* of the assessment. The factors that build the *task assessment* include the *marking criteria*, the *marking methods*, and the *skills and knowledge of the assessors*. The *marking criteria* for an assessment could be in the form of *marking rubrics*, *schemes*, *guides,* and *keys*. The *marking methods* used in this study were the *analytic marking* using a *marking rubric* which consisted of a set of criteria and the *comparative pairs judgements* using a holistic criterion.

Assessment quality indicates the level of confidence the stakeholders could have over the assessment result. The complete process of the assessment from the planning stage to the reporting should be designed to ensure that the assessment is *valid, reliable, authentic, transparent, equitable*, and *accountable*. In this study, because the focus was on the scoring method, the analysis was limited to the reliability of the scores and the validity of the scoring method. The reliability of the scores represents the consistency of the judgement process. The validity of the scoring method ensures that the scoring method measures what the assessment is aimed to measure.

Summary

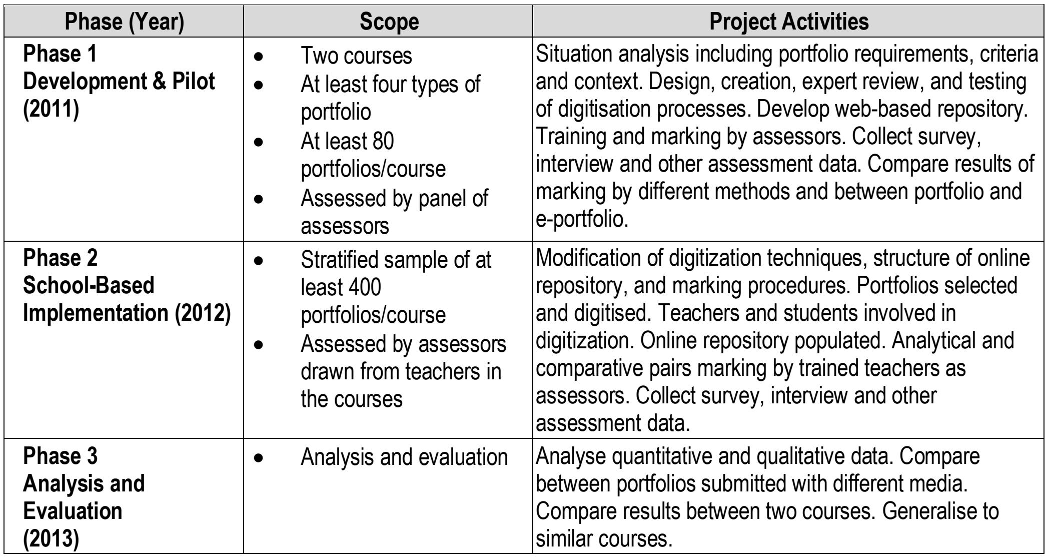
This chapter has discussed the main theoretical underpinnings related to this study. This study was grounded on two interconnected areas of theory: constructivism and of educational measurement, in particular Rasch models. These two areas of theory provided the foundation upon which the theories surrounding educational assessment discussed in this chapter was based. As such, a brief review of constructivism and educational measurement was presented, followed with discussions on the theories that stemmed from educational assessment that were pertinent to this study. These theories included those related to digital assessment, assessment task, task assessment, and assessment quality. This chapter was concluded with the conceptual framework of this study. The following chapter discusses the research methodology employed in the present study.

CHAPTER 3
RESEARCH METHODOLOGY

This chapter discusses the research methodology used in this study. The discussion includes the background in terms of the context and scope of this study, research design and the rationale, the population and sample, the various data collected, and the data analysis framework.

Background

The present study was conducted within the first phase of an ARC Linkage Project titled *The Authentic Digital Representation of Creative Works in Education*. This project was a collaborative research project between the Centre for Schooling and Learning Technologies (CSaLT) of Edith Cowan University and the Curriculum Council of Western Australia. Two secondary school subjects, Design and Visual Arts, were investigated. From this point onward, this overarching project would be called *the main project*. The main project was a four-year project that was divided into three phases. Figure 3.1 describes the three phases of the project (Newhouse, 2011a, p. 11).



Phase (Year)	Scope	Project Activities
Phase 1 **Development & Pilot** **(2011)**	• Two courses • At least four types of portfolio • At least 80 portfolios/course • Assessed by panel of assessors	Situation analysis including portfolio requirements, criteria and context. Design, creation, expert review, and testing of digitisation processes. Develop web-based repository. Training and marking by assessors. Collect survey, interview and other assessment data. Compare results of marking by different methods and between portfolio and e-portfolio.
Phase 2 **School-Based Implementation (2012)**	• Stratified sample of at least 400 portfolios/course • Assessed by assessors drawn from teachers in the courses	Modification of digitization techniques, structure of online repository, and marking procedures. Portfolios selected and digitised. Teachers and students involved in digitization. Online repository populated. Analytical and comparative pairs marking by trained teachers as assessors. Collect survey, interview and other assessment data.
Phase 3 **Analysis and Evaluation** **(2013)**	• Analysis and evaluation	Analyse quantitative and qualitative data. Compare between portfolios submitted with different media. Compare results between two courses. Generalise to similar courses.

Figure 3.1 Three phases of the Authentic Digital Representation of Creative Works in Education (Newhouse, 2011a, p. 11).

The present study was part of the first phase of the main project, which was the Development & Pilot phase. The main aim of the present study was to examine the suitability of the Comparative Pairs judgements as an alternative method of scoring for high-stakes practical assessment (Heldsinger & Humphry, 2010; Kimbell, 2008; Pollitt, 2012b; Pollitt & Whitehouse, 2012). This study adopted a mixed research method, using quantitative and qualitative data that were relevant to the purpose. These data included scores from three sources, assessors' notes, and interviews with the assessors. These were part of the data collected in the main project.

Context

This study was conducted in 2011 in Western Australia. As was outlined by the Curriculum Council of Western Australia, the courses for senior secondary school students were offered in four stages, which were Preliminary (P), Stage 1, Stage 2 and Stage 3, for each unit within the courses. The Preliminary Stage and Stage 1 units were designed as practical units to prepare students for either employment or future study, while Stage 2 and Stage 3 units were designed to prepare students for employment or further studies including in a university.

In 2011, as a requirement for the completion of secondary schooling, year 12 students in Western Australia who had been studying Stage 2 and Stage 3 courses undertook the WACE examinations. This examination was conducted by the Western Australian Curriculum Council and was applicable to all Western Australian students. As a high-stakes summative assessment, the WACE examination results provided information for students, teachers, tertiary education providers, employers, the government, and the general public. Among the many high-stakes purposes, this examination also provided "information to students about their achievement in a course to assist them in making decisions about post-school pathways" (Curriculum Council of Western Australia, 2011e, p. 62), as well as for tertiary education providers to assist with student placement. As such, Andrich (2006) emphasised the importance of assessments such as this "meet the requirements of being sufficiently rigorous and sufficiently *fine-grained* that they can be used for equitable selection into tertiary programs of study" (p. 3).

Each of the two courses investigated had a major practical component that contributed to 50% of the total WACE score. The nature of this practical component, however, was different. The Design course practical task was a 15-page portfolio that consisted of evidence of up to three Design projects. This evidence could be sketches, pictures, descriptions, and others. In Visual Arts this practical component was a finished artwork that could be two-dimensional, three-dimensional, or motion and time-based. The students were required to submit their finished artwork with an artist statement and installation pictures. It was anticipated that differences between the practical tasks in Design and Visual Arts potentially may contribute to differences in the scoring processes, issues surrounding the processes, and the scoring results themselves. The analysis of these variables was expected to highlight the characteristics of the Comparative Pairs judgements method in two different practical tasks. The results would contribute to the understanding of the way the Comparative Pairs judgements could be used in other subjects and for other purposes.

Scope

Aiming on investigating “the efficacy of digitisation and paired-comparisons method of judging of portfolios for the purposes of summative assessment in the Visual Arts and Design senior secondary school courses” (Newhouse, Pagram, Paris, Hackling, & Ure, 2012, p. 8), the main project delved into a broader theme than this study. *Figure 3.2* illustrates the position of this study inside the main project. From the four data sources obtained in the main project; survey data from the students who were involved, interview with their teachers, data from the scoring processes and interview with the assessors; only the latter two sources of data were used for this study. Because this study was focussed on the quality of the result from Comparative Pairs judgements method, only the scoring data and the assessor interview were considered pertinent.

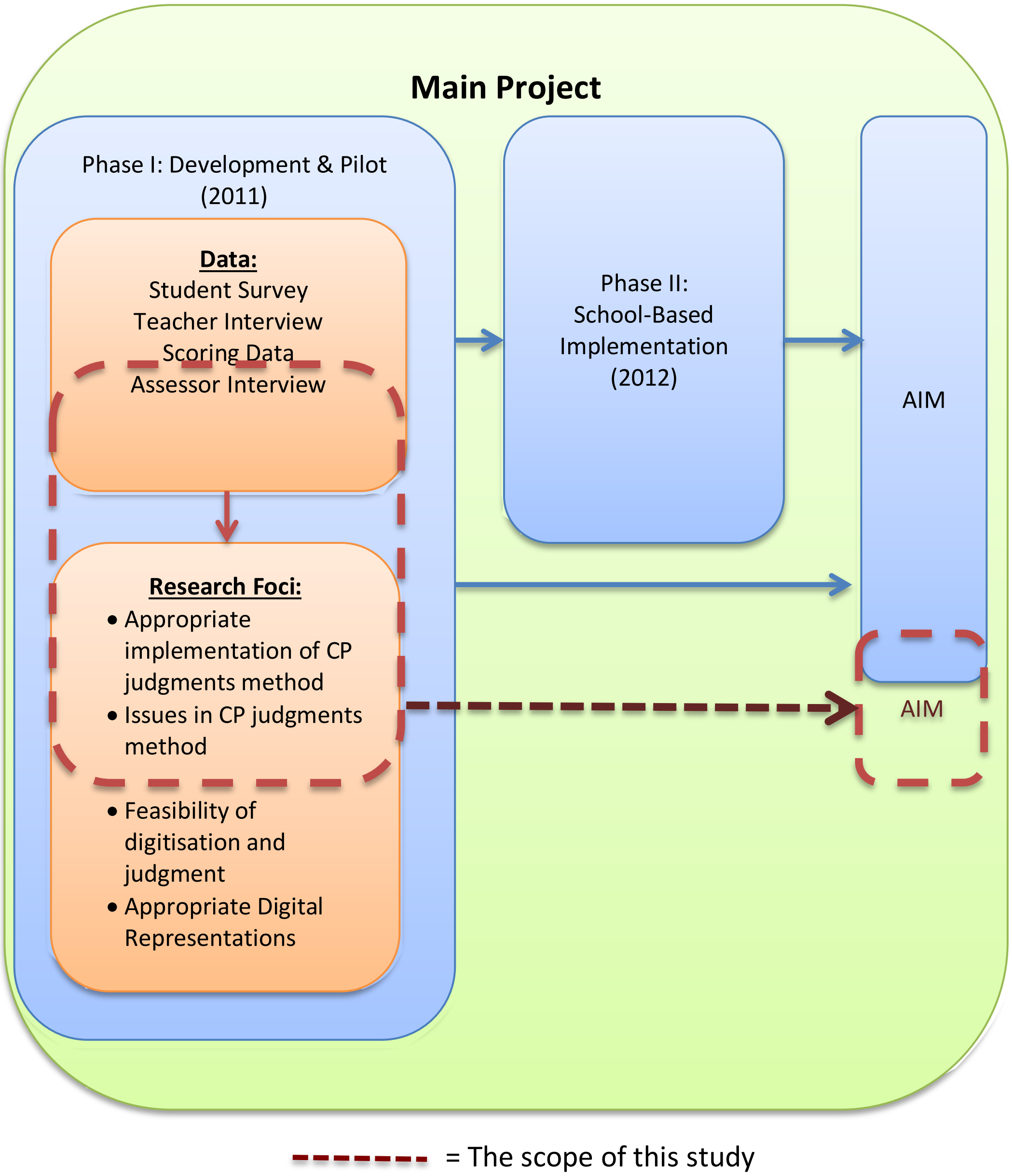

Main Project
Phase I: Development & Pilot (2011)
Data:
Student Survey
Teacher Interview
Scoring Data
Assessor Interview
Research Foci:
Appropriate implementation of CP judgments method
Issues in CP judgments method
Feasibility of digitisation and judgment
Appropriate Digital Representations
Phase II: School-Based Implementation (2012)
AIM
AIM
= The scope of this study

Figure 3.2 The scope of this study.

Research Methodology

This study borrowed from the interpretive research paradigms (Assalahi, 2015) by utilising mixed research methods in case studies and cross case analysis. Assalahi argues that an educational research "paradigm consists of at least three elements; ontology, epistemology and methodology" (p. 313). These three elements (or constructs) should align and support each other. For example, a positivistic paradigm typically has objectivism as an ontology, realism as an epistemology, and empiricism as a methodology. A positivist paradigm tends to be associated with research in the natural sciences, although is often referred to as a normative paradigm in the social sciences (Cohen, Manion, & Morrison, 2011).

The current study follows a more interpretive paradigm that was developed in opposition to positivism (Assalahi, 2015) or the normative paradigm (Cohen et al., 2011). Interpretivist researchers consider that human behaviours are set within interactions in a world context that is highly subjective. Therefore the ontological stance is subjectivism or relativism with a social constructivist epistemology (Assalahi, 2015). These were discussed in Chapter Two as the theoretical framework that guided the review of the literature. In particular this suggested an investigation of the literature on authentic assessment and educational measurement. To align with these elements of the interpretive paradigm the theoretical framework for the methodology and research design is fundamentally qualitative in nature (Assalahi, 2015).

Assalahi (2015) suggests a number of interpretive approaches to educational research: ethnography; phenomenology; and case study. While the current study could be considered to include aspects of each of these, it is more accurately aligned with a case study approach. The two courses, Design and Visual Arts, were treated as separate case studies as it was recognised that the curriculum, students and teachers have different characteristics in each so that each could be considered an "occurrence" (p. 315). The aim was to uncover "the reasons behind the occurrence of a thing" and to discern the "interrelated factors" (p. 315).

This type of qualitative research seeks to gain a "deeper understanding, by means of collecting and categorizing, of data and actions of participants … rather than generalizing" (Assalahi, 2015, p. 315). Thus a consideration of the perceptions of assessors was an important component of the analysis of the data. Assessors brought their own construction of knowledge concerning assessment and what represented a good performance by a student. But equally the study considered an understanding that the portfolios being assessed were expressions of a social reality of the students creating them and that assessors interacted with their perception of this reality based on their previous experience with students. Thus the data included not only interviews with assessors and records of their notes made while assessing but also investigated misfit statistics both in terms of the assessors and the portfolios.

Research method

This study employed the *explanatory sequential mixed methods design* (Creswell, 2008), a mixed methods research design which focuses mainly on the quantitative data and utilises the qualitative data to explain the phenomenon investigated in the research. Only the follow-up explanations variant (Creswell & Plano Clark, 2011) of this research design was used.

The aim of this study was to investigate the suitability of the Comparative Pairs judgements method in assessing practical tasks in two subjective courses: Design and Visual Arts. The suitability of this scoring method was examined using the reliability of the scores and the validity of this method. The explanatory sequential mixed methods design as is shown in *Figure 3.3* was considered to be the best research design for this purpose. (Creswell, 2008) described this design as "captures the best of both quantitative and qualitative data – to obtain quantitative results from a population in the first phase, and then refine or elaborate these findings through an in-depth qualitative exploration in the second phase" (p. 560). He went on to explicate that the problematic part of this design was in deciding which components to explore further.

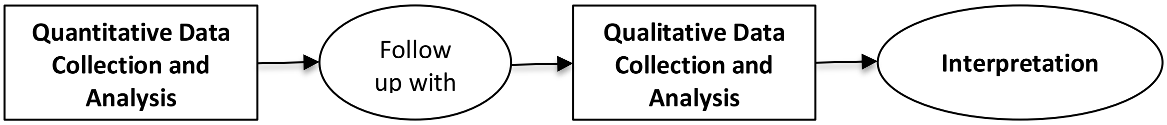

Quantitative Data Collection and Analysis
Follow up with
Qualitative Data Collection and Analysis
Interpretation

Figure 3.3 Explanatory Sequential Design (Creswell, 2012, p. 542).

Creswell and Plano Clark (2011) introduced two ways to determine the selection of these components. The first was the *follow-up explanations* variant, which uses the qualitative data to find possible explanations for the quantitative data. The second was the *participant selection* variant, which focussed the data analysis more on the qualitative data. In this study, only the first variant was used because it fit the research questions of this study.

The main data source in this study was the results from the scoring processes. These quantitative data were analysed to produce the necessary descriptive data and to statistically test the characteristics of the scoring methods, especially the Comparative Pairs judgements method. There were three scoring processes that provided the scoring data: the WACE practical examination, the online Analytical marking, and the Comparative Pairs judgements method. The scoring results were analysed using SPSS statistics software and RUMM, a Rasch modelling software. The supplementary qualitative data in this study were obtained from the assessor interview data and the assessors' notes recorded in the two scoring systems that were used in the scoring processes.

Aligned with the research questions for this study, the main purpose of the data analysis was to examine the reliability of the scoring results and the validity of the Comparative Pairs judgements method. The data analysis process started with descriptive statistics of the scoring results to provide a general description of the scores. This analysis was followed by the reliability and validity analysis of the scores obtained from the Comparative Pairs judgement. A discrepancy analysis was conducted afterwards to investigate the possible patterns and explanations for portfolios that are scored too differently by different assessors or in different scoring methods. In this step the follow-up explanations variant of the explanatory research design (Creswell & Plano Clark, 2011) was used to examine the factors that might affect the discrepancy, using data from the assessor interview and assessors' notes from the scoring systems.

Population and samples

In 2011, 403 students undertook the Stage 3 Design WACE examination across four Design contexts, which were Photography, Graphics, Dimensional Design, and Technical Graphics. In Visual Arts course there were 926 Stage 3 WACE examination students. Together with the Curriculum Council of Western Australia, researchers from the project chose the schools and teachers in the Perth Metro area to be invited to participate in the project.

Students from schools that were willing to participate were then given an information letter and a consent form to be read and signed by their parents or guardians. In the Design course, six teachers and 82 students agreed to participate, while in Visual Arts there were ten teachers and 75 students. Table 3.1 describes the participating schools and students.

Table 3.1
List of Schools and Students Involved in the Study

Case	School Type	Number of Classes	Number of Students*	Course	Context
VC	Private – Co Ed	1	3	Visual Arts	Varied
VH	Private – Co Ed	1	3	Visual Arts	Varied
VJ	Private – Co Ed	1	10	Visual Arts	Varied
VK	Public	1	9	Visual Arts	Varied
VL	Public	1	11	Visual Arts	Varied
VN	Public	1	10	Visual Arts	Varied
VO	Private	1	4	Visual Arts	Varied
VP	Private	1	7	Visual Arts	Varied
VQ	Private	1	5	Visual Arts	Varied
VS	Private	1	13	Visual Arts	Varied
DB	Public	1	4	Design	Photography
DL	Public	1	18	Design	Photography
DM	Private	1	13	Design	Technical graphics
DN	Public	1	17	Design	Photography
DT	Private	1	21	Design	Technical graphics
DV	Public	1	9	Design	Graphics

* Number of students consenting to be involved with the study, not the number in the class.

Data Collection and Analysis

This section discusses the collection of data in this study, and the research method to analyse those data. Data collected included the scoring data and assessors' notes obtained from the two online scoring processes, scoring data from the WACE practical examination, and interview with the assessors from the two online scoring methods.

In reference to the conceptual framework in Chapter 2, the data collection process in this section is also discussed in two parts. The first part discusses the Assessment Task to provide the background information on the type of assessment task involved in each course. The second part of the discussion is on the Task Assessment, which is a discussion on the scoring methods that were used to obtain the scoring data used in this study.

Assessment task - Design

In Western Australia, the Design course consisted of four specific course contexts which were photography, graphics, dimensional design and technical graphics (Curriculum Council of Western Australia, 2010a). In each course the context of the program was focussed on three areas of course content; which were design principles and process, communication principles and visual literacies, and production knowledge and skills; aside from the context-specific skills. At the end of year 12, the Design students sat the Western Australian Certificate of Education (WACE) examination. The 2011 WACE examination consisted of two components, written and practical, each contributed to 50% of the total score. The written examination was divided into two sections, short response and extended response. The practical examination was in the form of a Design portfolio that consisted of examples of the development of two or three Design projects on which they had been working. Details on the task are discussed in Chapter 4.

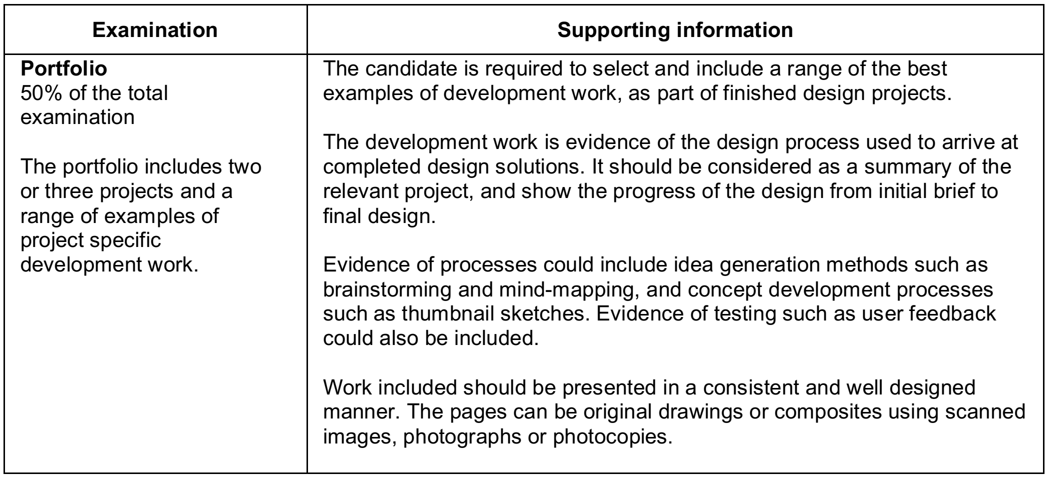
The marking of student practical work was aimed to provide a “fair and equitable ranking” (Curriculum Council of Western Australia, 2011a, p. 3). To facilitate the marking process, the Curriculum Council developed a marking rubric and a marking guideline to be used by the markers. They also conducted briefings, meetings and trainings for the markers. The WACE Design marking rubric can be viewed in Appendix C. There were two markers for each Design portfolio to increase objective and fair marking, under the supervision of a Chief Marker whose role included mediation in the event of no agreed mark was reached between the two markers.

These steps were taken by the Curriculum Council to ensure the “accuracy, fairness and manageability” of the marking process (Curriculum Council of Western Australia, 2011a, p. 3). Even though these steps were taken and research studies indicated that these steps could increase the accuracy, and with accuracy, fairness, the subjective nature and the complexity of the practical assessment were still likely to affect the accuracy of the assessment result. Aside from the potential threats to accuracy associated with the use of a marking rubric as was discussed in the previous section, the issues in a Design course according to Kimbell (2007) could also stem from the many various possible ways that students could solve a problem in Design. Several Design course educators, therefore, considered using holistic assessment for the course (Wooff, Bell, & Owen-Jackson, 2013).

For the practical component of the WACE examination Design students were required to submit a Design portfolio that represented their understanding and practical skills in the production of design. The portfolio had to include (Curriculum Council of Western Australia, 2010a; Newhouse et al., 2012):

- an index of the contents identifying each project;
- a checklist that indicates all documents conform to portfolio specifications;
- the completed Designer statement;
- the completed References/acknowledgement form; and
- the design project (15 A3 pages).

The Design syllabus (Curriculum Council of Western Australia, 2010a, p. 35) provided a guideline for the submission of the portfolio as follows:



Examination	Supporting information
Portfolio 50% of the total examination The portfolio includes two or three projects and a range of examples of project specific development work.	The candidate is required to select and include a range of the best examples of development work, as part of finished design projects. The development work is evidence of the design process used to arrive at completed design solutions. It should be considered as a summary of the relevant project, and show the progress of the design from initial brief to final design. Evidence of processes could include idea generation methods such as brainstorming and mind-mapping, and concept development processes such as thumbnail sketches. Evidence of testing such as user feedback could also be included. Work included should be presented in a consistent and well designed manner. The pages can be original drawings or composites using scanned images, photographs or photocopies.

Figure 3.4 Design practical (portfolio) examination design brief (Curriculum Council of Western Australia, 2011a, p. 35).

WACE practical examination in Design aimed to show student achievement in the Design course. For this task, Design students were required to provide evidence of the design processes that lead to the realisation of up to three final design projects in the portfolio. This evidence could include:

- brainstorming and mindmapping as idea generation methods
- analysis of information and translation into design concepts
- application of design principles
- visualisation of concepts
- application of interrelated thinking and innovative development process such as thumbnail sketches
- use of interpretive skills and problem solving
- selection and use of a diverse range of skills, techniques and procedures
- application of planning and production methods
- use of design elements
- evidence of testing such as user feedback.

The variety of forms that the Design students had included:

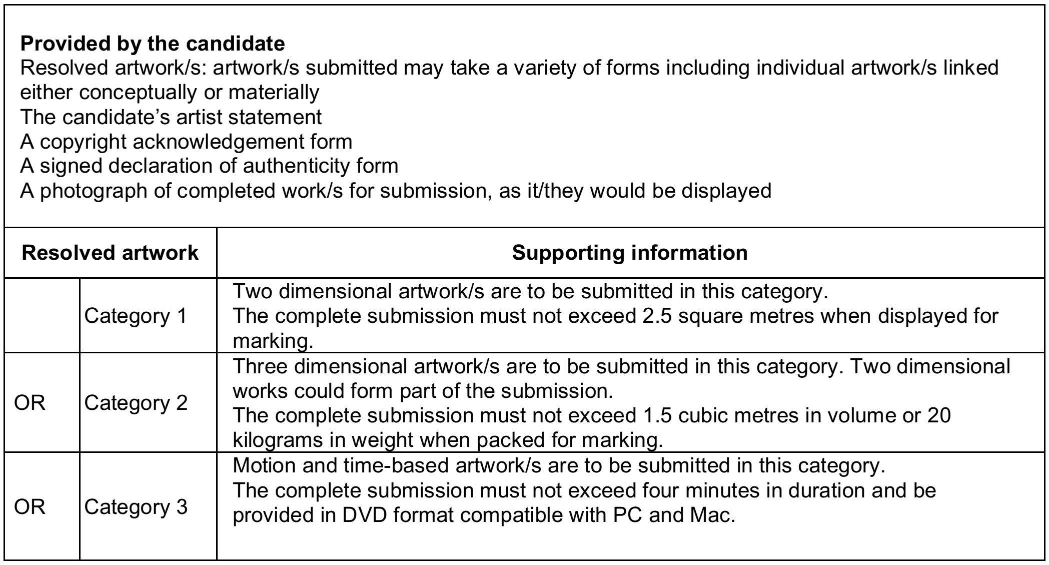
- a series of design projects in one genre or style
- works that are linked either conceptually or materially
- individual design projects that employ a variety of production methods.

Besides serving as a practical examination work, this Design portfolio were also used as a cheat sheet for the written examination that the students could bring along into the written examination room.

Assessment task - Visual Arts

In Western Australia, the Visual Arts course was among the examination courses for the the Western Australian Certificate of Education (WACE). The Visual Arts course valued “divergence, uniqueness and individuality” (Curriculum Council of Western Australia, 2010b, p. 3). The course was focussed on building the students’ knowledge, skills, understanding and appreciation in visual arts, and their ability to solve problems in an innovative and creative way. There were two content areas in this course. The areas were *art making*, which included the understanding and skills in techniques and processes, and *art interpretation*, which included arts analysis and critique.

As in the Design course, the WACE examination for the Visual Arts course also consisted of a written examination and a practical examination, each of which contributed for 50% of the total WACE score. The written examination consisted of three sections, which were the short answer section, the compare and contrast essay and the investigation essay, each of which contributed to 10%, 15% and 25% consequently. The practical examination in Visual Arts was in the form of a finished artwork in either of a two-dimensional artwork, a three-dimensional artwork, or a motion and time-based artwork. The detailed requirement of the practical examination submission can be viewed in Appendix B. This artwork was to be accompanied by an artist statement that provides the background information on the artwork, however, this statement was not to be scored. The marking process of Visual Arts WACE examination was similar to that of the Design course. The marking key can be viewed in Appendix D. The Visual Arts syllabus (Curriculum Council of Western Australia, 2010b) provided a guideline for the submission of the artwork as follows:



Provided by the candidate Resolved artwork/s: artwork/s submitted may take a variety of forms including individual artwork/s linked either conceptually or materially The candidate's artist statement A copyright acknowledgement form A signed declaration of authenticity form A photograph of completed work/s for submission, as it/they would be displayed		
Resolved artwork		**Supporting information**
	Category 1	Two dimensional artwork/s are to be submitted in this category. The complete submission must not exceed 2.5 square metres when displayed for marking.
OR	Category 2	Three dimensional artwork/s are to be submitted in this category. Two dimensional works could form part of the submission. The complete submission must not exceed 1.5 cubic metres in volume or 20 kilograms in weight when packed for marking.
OR	Category 3	Motion and time-based artwork/s are to be submitted in this category. The complete submission must not exceed four minutes in duration and be provided in DVD format compatible with PC and Mac.

Figure 3.5 Visual Arts practical (production) examination design brief (Curriculum Council of Western Australia, 2010b, p. 33).

As mentioned in *Figure 3.5*, the students were required to submit their exhibition-ready artwork that could be a single work, or a collection or a suite of single works that were conceptually or materially linked. Together with their artwork, the students needed also to include a few documents, including an artist statement. The artist statement was a written summary that explained the student's thinking process until the realisation of the artwork that could include the original source of the idea, the significance of the artwork, an explanation on the material or technique used and other information that might be needed to accompany the artwork.

The most distinguishable difference between the assessment tasks in Design and Visual Arts was the type of the task. In Design the assessment task was in the form of a portfolio that displayed student developmental work while in Visual Arts the task was in the form of a finished artwork. Messick (1994) categorised these different tasks as *performance-and-product assessment*. In Design the task showed the process of students' design development while in Visual Arts the task showed a final product, with only an artist statement to give a brief information on the students' creation process. This difference in the type of the assessment task and how it might affect the scoring results is discussed in Chapter 6.

Digitisation and scoring of student work

This study was conducted within the first phase of the main project, which was the pilot phase of the project. In this first phase the researchers involved in the project digitised the student work. The digitisation process and the issues that arose from different factors of the digitisation process were analysed and used to develop a set of guidelines to help the students participating in the second phase digitise their work for online submission.

Design

The digitisation process of student design portfolios was done on site in one of the buildings used by the Curriculum Council for storing and marking the Design portfolios. The portfolios were scanned and combined to produce one pdf file that was saved in each student's folder that was named according to their assigned ID. Each student had a pdf portfolio that consisted of the 15 pages and the required addenda, which included an index, a checklist, a designer statement, and a references/acknowledgement sections.

For the Analytical Marking, the students' pdf files were saved on an ECU web server, together with the Filemaker Pro database that was developed for this purpose. *Figure 3.6* describes the digitisation and scoring process for Design course.

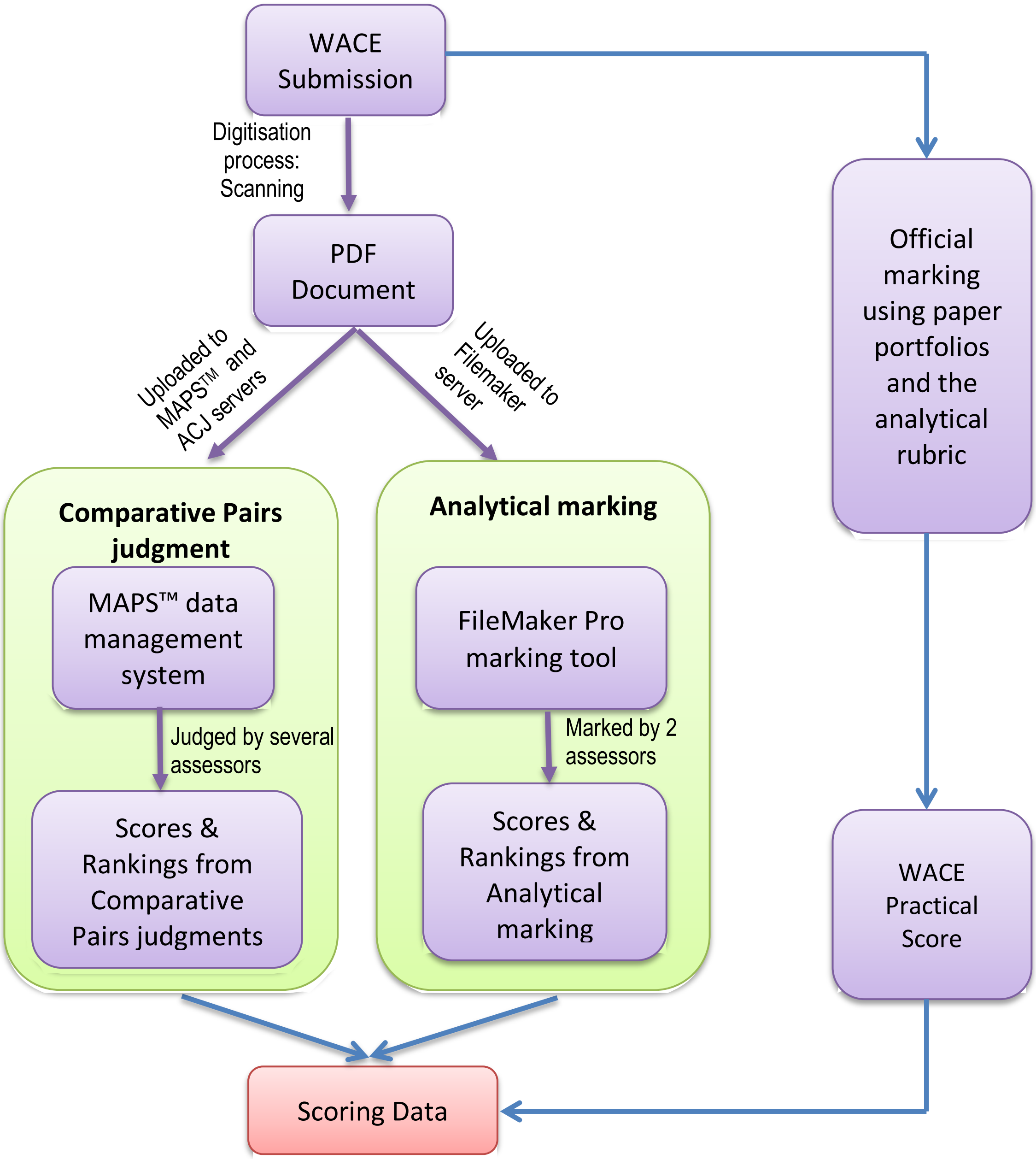

WACE Submission
Digitisation process: Scanning
PDF Document
Uploaded to MAPS™ and ACJ servers
Uploaded to Filemaker server
Official marking using paper portfolios and the analytical rubric
Comparative Pairs judgment
MAPS™ data management system
Judged by several assessors
Scores & Rankings from Comparative Pairs judgments
Analytical marking
FileMaker Pro marking tool
Marked by 2 assessors
Scores & Rankings from Analytical marking
WACE Practical Score
Scoring Data

Figure 3.6 Digitisation and scoring process for the Design course.

Visual Arts

The digitisation process of Visual Arts work was done on site in the display room for the WACE examination marking. In the digitisation process the artworks were assembled according to the accompanying installation picture and digitised. The resulting digital representation of the artworks included photographs from different angles and videos. One photograph was taken for two-dimensional artworks and five were taken for three-dimensional artworks, each featuring the front, right, back, left and top side. For two-dimensional artworks, only a video panning from side-to-side and zooming into the artwork was taken. For small three-dimensional artworks, a round table was used to video all the angles on the artworks. For larger three-dimensional artworks, for which the use of the round table was not possible, a manual video capture was taken by moving the video around the artworks.

An experienced Visual Arts educator helped with choosing four close-up captures from each student's set of photographs. The purpose of these close-ups was to provide the examiners with a detailed view on several particular parts of the artworks. These close-ups highlighted factors such as textures and layers.

Because the resolution of the digital SLR cameras was very high, the size of the original digital photographs was too big to be used in online scoring processes. Therefore, the digital files needed to be resized. The approved file size was 72 dpi. In this size the files were sufficient for online scoring without compromising the visual presentation of the photographs.

For the online marking processes, a PowerPoint file was prepared for each student, consisting of:

- artist statement
- student's installation photograph
- resized photographs
- close-ups.

This PowerPoint file was subsequently converted into a pdf file to prevent inadvertent changes during scoring. During the scoring processes, the assessors had an access to the students' folders which each consisted of the pdf file, the video files, and a folder that contained the individual files that were also already compiled into the pdf file. Table 3.2 shows a list of these files.

Table 3.2
Visual Arts Portfolio Files

Filename	Description
as.jpg	**Artist statement**
	Student's description of the artwork which could include the original idea, chosen artwork media, or techniques.
c1.jpg	**Close-ups**
(c1, c2, c3, c4)	Cropped photographs to highlight certain qualities of the artwork.
Install.jpg	**Student's installation picture**
	Student's photograph of the artwork the way they would like to install it as for an exhibition.
p1.jpg	**Full photographs**
(p1, p2, p3, p4, p5 for 3D)	Photographs of the artwork that had been resized to 72 dpi.
ppt.pptx	**PowerPoint**
	PowerPoint file containing the full set of photographs.
pdf.pdf	**PDF file**
	The PowerPoint file that had been converted into a PDF file.
v.mov (vr.mov for 3D)	**Quicktime video file**
	Video of the artwork formatted for Mac computers.
v.wmv (vr.wmv for 3D)	**Windows media file**
	Video of the artwork formatted for Mac computers.

The portfolios were then uploaded into the servers for the two online scoring processes. For the Comparative Pairs judgements, the portfolios were made available online to assessors through the ACJ system installed on a Curriculum Council server. For the Analytical marking, the portfolios were uploaded into an ECU server and were accessible through usernames and passwords on the FileMaker Pro marking tool that was developed specifically for this process. *Figure 3.7* depicts the complete digitisation process for Visual Arts.

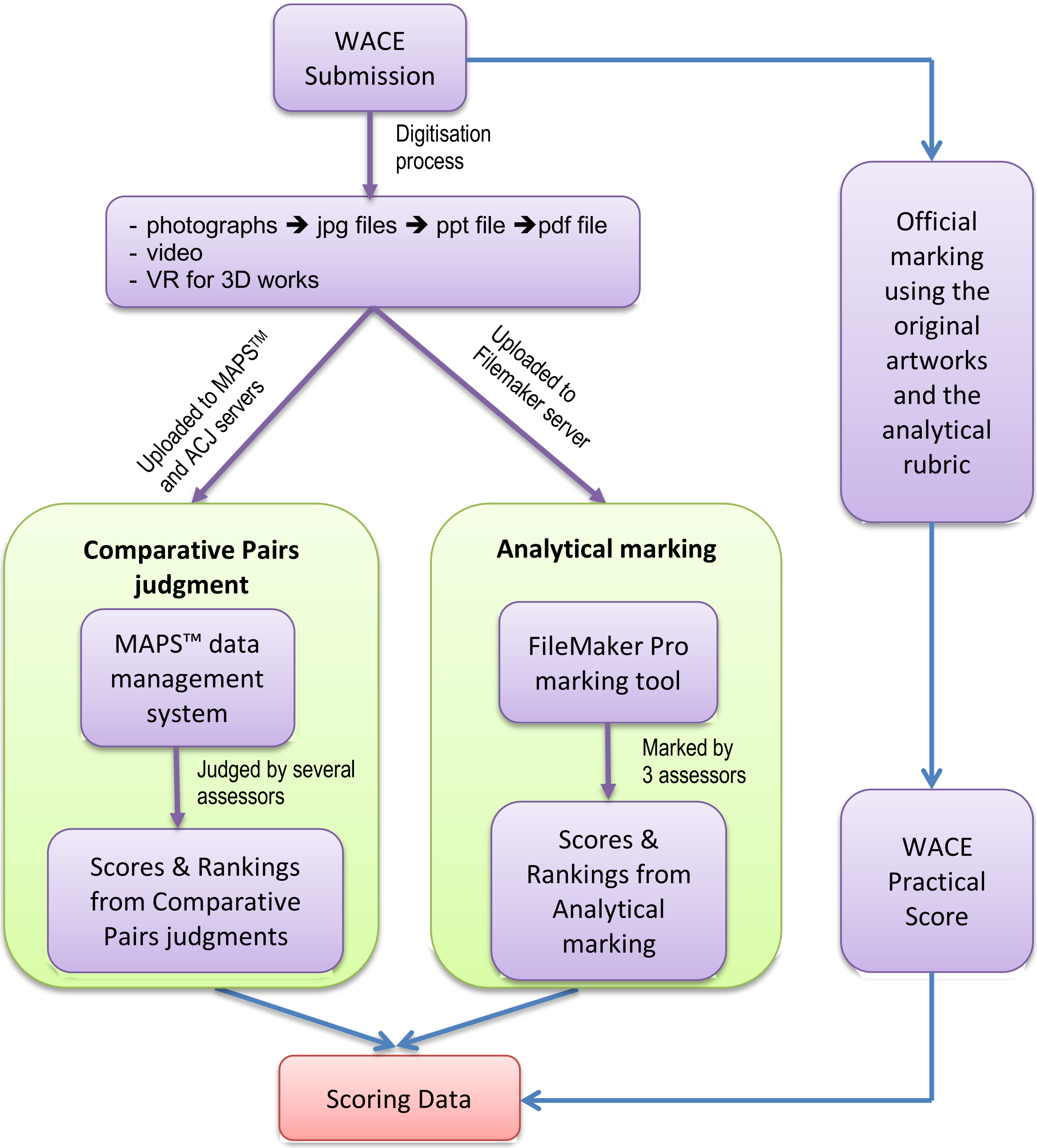

WACE Submission
Digitisation process
- photographs ➔ jpg files ➔ ppt file ➔pdf file
- video
- VR for 3D works
Uploaded to MAPS™ and ACJ servers
Uploaded to Filemaker server
Official marking using the original artworks and the analytical rubric
Comparative Pairs judgment
MAPS™ data management system
Judged by several assessors
Scores & Rankings from Comparative Pairs judgments
Analytical marking
FileMaker Pro marking tool
Marked by 3 assessors
Scores & Rankings from Analytical marking
WACE Practical Score
Scoring Data

Figure 3.7 Digitisation and scoring process for Visual Arts course.

Task assessment

Scoring data for this study were obtained through three scoring processes: the Comparative Pairs judgements, the Analytical marking and the WACE practical examination marking. The two first-mentioned processes were conducted online within the project, while the WACE marking was conducted by the Curriculum Council. This section discusses these processes.

Comparative Pairs judgement

The Comparative Pairs judgements process was started in two separate half-day workshops at Edith Cowan University campus, one for each course. The aim of the workshops was to develop a holistic criterion for the scoring and to introduce the assessors to the Adaptive Comparative Judgement (ACJ) system, the *pairs engine* that was used for the Comparative Pairs judgements process. Based on the same set of criteria used in the rubric for the Analytical marking, the assessors discussed and then decided on one holistic criterion that they considered could best represent the criteria. This holistic criterion would then be used by the assessors for the Comparative Pairs judgements.

The Adaptive Comparative Judgement system was the online scoring system that was used in the Comparative Pairs judgements process in the project and this study. The system was developed within the Technology Education Research Unit (TERU) project at the Goldsmiths, University of London. It was a collaboration project of Alistair Pollitt, TAG Learning, and Goldsmith College (Kimbell et al., 2009). This web-based assessment system was an integrated assessment system that managed student digital work repository, created pairings of student work, displayed the pairings for judgement, and provided a statistical analysis of the result from the judgement process. This system has been trialled and used in several institutions in several countries such as UK, Singapore, Sweden and Spain.

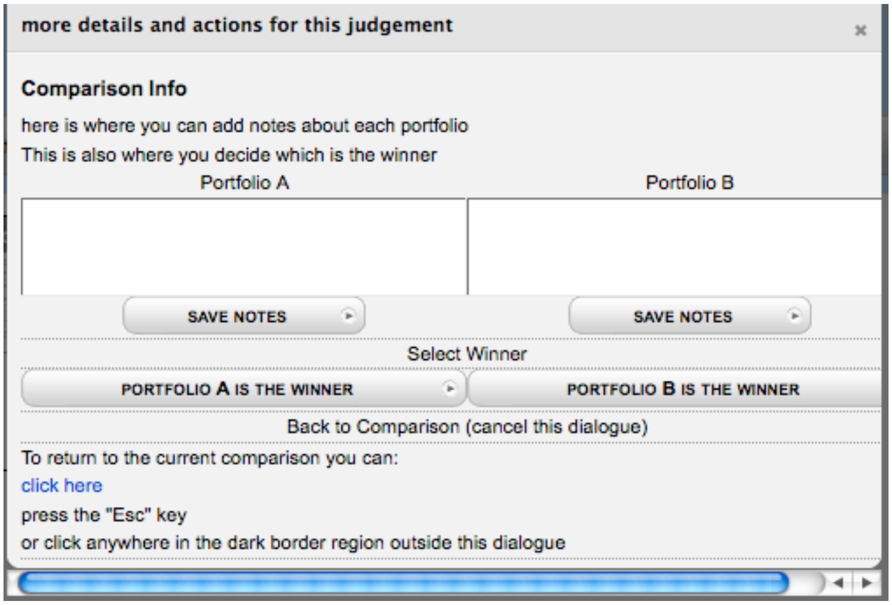
In this study, the student work in the two courses, Design and Visual Arts, was digitised and uploaded into SCaSA server with structures that were described in *Figure 3.6* and *Figure 3.7*. The ACJ system then created the pairings randomly for the first round of judgements. In this first round, the scoring process resulted in 50% *winners* and 50% *losers*. In the second round the pairs engine paired works 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 continued on to create pairings that "will provide the most information for increasing the reliability of the rank order" (Pollitt & Whitehouse, 2012, p. 4) by matching pairs of work that were of more and more similar quality. Because of this adaptive function this system is called the Adaptive Comparative Judgement system (Pollitt, 2012a; Pollitt & Whitehouse, 2012).

Starting from the seventh round, a different pairing method was 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 assessors and increase the efficiency of the judging process (Kimbell, 2008). If after all of the works were paired up the reliability coefficient was still considered not sufficiently high, another round was created. Once a reasonably high reliability had been achieved, the scoring process was stopped and the scoring data were processed to be analysed.

At the end of each round, the pairs engine could provide information about the judging sessions, including the reliability achieved up to that point. At the end of the process, when a high reliability has been attained, the ACJ system provided data that described judgement misfits, which included agreement among assessors, assessors' notes for each pair of work, assessors' notes for each judgement, the Rasch parameter location of each work, and the time needed for the judgement. The flexibility of the system allowed for inconsistency in the number of assessors making judgement in each judgement session and the variety in the number of judgements among assessors.

During the judgement process, the assessor logged onto the ACJ system using their unique ID and password. A screenshot of the judgement page for the Design course is shown in *Figure 3.8*. Once they were logged into the system, they could start their judgement session. For the judgement, a pair of portfolios would be displayed for them in thumbnail images that could be expanded, as is shown in *Figure 3.9*. The assessors make their judgements by clicking on either *PORTFOLIO A IS THE WINNER* or *PORTFOLIO B IS THE WINNER* button. The *Comparison Info* boxes were where the assessor could type in their judgement notes, which were the notes about the comparison. Portfolio A and Portfolio B notes boxes were the assessors' notes for each portfolio. These notes would be attached to each particular portfolio and viewable only to that particular assessor to help him when that portfolio was again part of another pair he needed to judge later. After one judgement is finished, the system presented the next pair for the assessor to judge.



more details and actions for this judgement

Comparison Info

here is where you can add notes about each portfolio

This is also where you decide which is the winner

Portfolio A

Portfolio B

Select Winner

Back to Comparison (cancel this dialogue)

To return to the current comparison you can:

click here

press the "Esc" key

or click anywhere in the dark border region outside this dialogue

Figure 3.8 ACJ judgement page for the Design course.

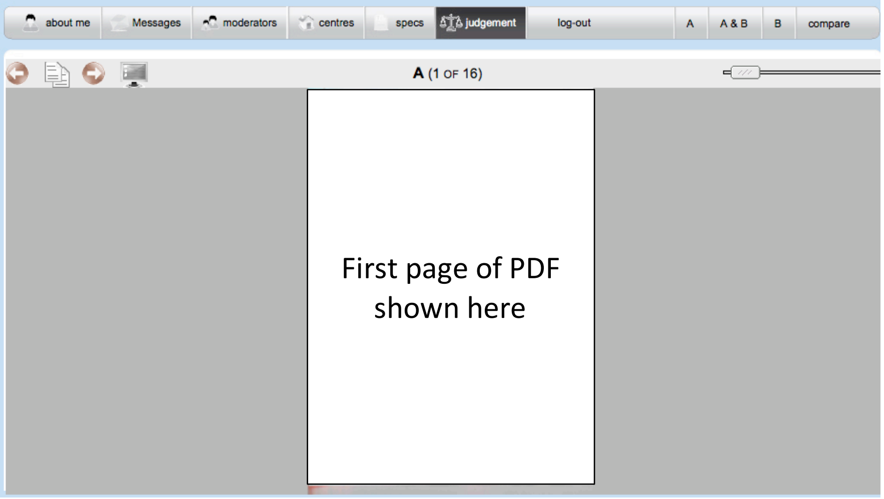


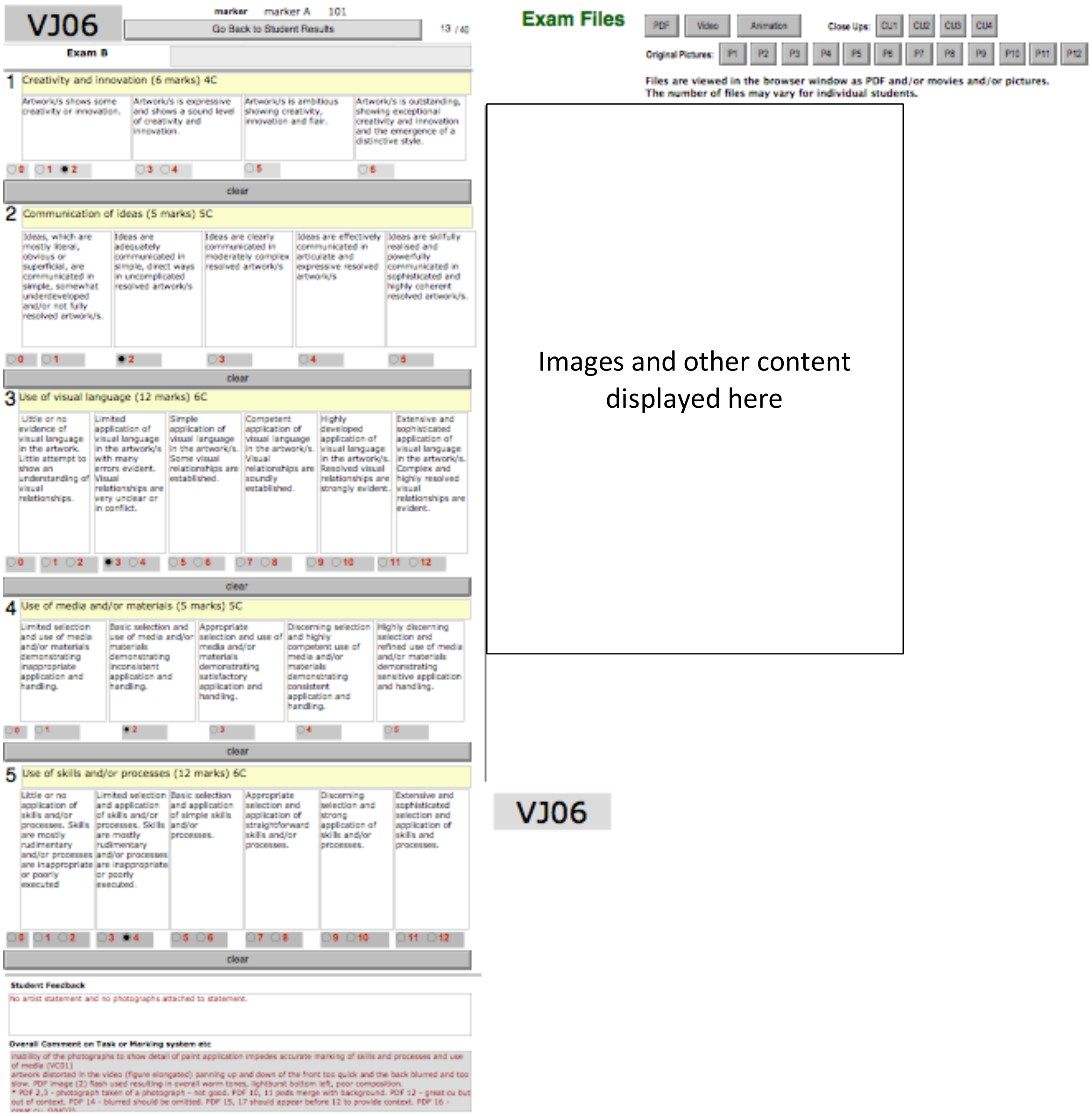
Figure 3.9 Expanded view of portfolio A.

Analytical marking

The Analytical marking process was supported by a Filemaker Pro database developed for the main project. Two experienced assessors in Design and three in Visual Arts were asked to analytically mark all Design portfolios. Considering all assessors were highly experienced in WACE marking, instead of a workshop, there was only a meeting to ensure there was a mutual understanding on the marking criteria and the assessors' ability to use the online marking interface. A rubric devised by the Curriculum Council was built into this database to create an online marking tool for each course. The assessors for each course marked each digital portfolio using this interface. The Curriculum Council marking rubric for the Design course practical examination in 2011 consisted of six criteria with various weightings. The total score for the practical component was 50. In Visual Arts, the marking rubric consisted of five criteria with different weightings. The total score for the practical component was 40. These analytical marking criteria could be viewed in Appendices C and D.

Assessors used a standard Internet browser to connect to the FileMaker Pro database. They were able to view the representations of student work and record their judgements as scores on the rubric. The assessors marked the student work on a marking page that consisted of two sections: the marking section and the viewer section. In the viewer section on the right-hand side of the marking page, the assessors could choose the type of file they needed to view. For the Design course there was only one button to show the student PDF file while for the Visual Arts course there were several buttons, one for each type of files.

The marking rubric was located on the left side of the marking page, with a button for every score underneath every outcome in each criterion to make the marking process quicker and easier. At the bottom of the marking page there were two comment boxes, one was for feedback for each of the student work while the other one was for the overall comment for the marking system and process. While the assessors were encouraged to use these comment boxes, only a few assessors did so. below shows the marking page for Visual Arts.



VJ06

marker marker A 101

13 / 40

Exam B

Files are viewed in the browser window as PDF and/or movies and/or pictures.
The number of files may vary for individual students.

1 Creativity and innovation (6 marks) 4C

Artwork/s shows some creativity or innovation.	Artwork/s is expressive and shows a sound level of creativity and innovation.	Artwork/s is ambitious showing creativity, innovation and flair.	Artwork/s is outstanding, showing exceptional creativity and innovation and the emergence of a distinctive style.
○ 0 ○ 1 ● 2	○ 3 ○ 4	○ 5	○ 6

2 Communication of ideas (5 marks) 5C

Ideas, which are mostly literal, obvious or superficial, are communicated in simple, somewhat underdeveloped and/or not fully resolved artwork/s.	Ideas are adequately communicated in simple, direct ways in uncomplicated resolved artwork/s	Ideas are clearly communicated in moderately complex resolved artwork/s	Ideas are effectively communicated in articulate and expressive resolved artwork/s	Ideas are skilfully realised and powerfully communicated in sophisticated and highly coherent resolved artwork/s.
○ 0 ○ 1	● 2	○ 3	○ 4	○ 5

3 Use of visual language (12 marks) 6C

Little or no evidence of visual language in the artwork. Little attempt to show an understanding of visual relationships.	Limited application of visual language in the artwork/s with many errors evident. Visual relationships are very unclear or in conflict.	Simple application of visual language in the artwork/s. Some visual relationships are established.	Competent application of visual language in the artwork/s. Visual relationships are soundly established.	Highly developed application of visual language in the artwork/s. Resolved visual relationships are strongly evident.	Extensive and sophisticated application of visual language in the artwork/s. Complex and highly resolved visual relationships are evident.
○ 0 ○ 1 ○ 2	● 3 ○ 4	○ 5 ○ 6	○ 7 ○ 8	○ 9 ○ 10	○ 11 ○ 12

4 Use of media and/or materials (5 marks) 5C

Limited selection and use of media and/or materials demonstrating inappropriate application and handling.	Basic selection and use of media and/or materials demonstrating inconsistent application and handling.	Appropriate selection and use of media and/or materials demonstrating satisfactory application and handling.	Discerning selection and highly competent use of media and/or materials demonstrating consistent application and handling.	Highly discerning selection and refined use of media and/or materials demonstrating sensitive application and handling.
○ 0 ○ 1	● 2	○ 3	○ 4	○ 5

5 Use of skills and/or processes (12 marks) 6C

Little or no application of skills and/or processes. Skills are mostly rudimentary and/or processes are inappropriate or poorly executed	Limited selection and application of skills and/or processes. Skills are mostly rudimentary and/or processes are inappropriate or poorly executed.	Basic selection and application of simple skills and/or processes.	Appropriate selection and application of straightforward skills and/or processes.	Discerning selection and strong application of skills and/or processes.	Extensive and sophisticated selection and application of skills and processes.
○ 0 ○ 1 ○ 2	○ 3 ● 4	○ 5 ○ 6	○ 7 ○ 8	○ 9 ○ 10	○ 11 ○ 12

Student Feedback

No artist statement and no photographs attached to statement.

Overall Comment on Task or Marking system etc

inability of the photographs to show detail of paint application impedes accurate marking of skills and processes and use of media (VC01)
artwork distorted in the video (figure elongated) panning up and down of the front too quick and the back blurred and too slow. PDF image (2) flash used resulting in overall warm tones, lightburst bottom left, poor composition.
* PDF 2,3 - photograph taken of a photograph - not good. PDF 10, 11 pods merge with background. PDF 12 - great cu but out of context. PDF 14 - blurred should be omitted. PDF 15, 17 should appear before 12 to provide context. PDF 16 - great cu. [illegible]

VJ06

Figure 3.10 Filemaker Pro analytical marking page.

Once an assessor finished marking a student work, the marking tool calculated the score for that student and saved the score in the database for data analysis. The assessor could also view his marking result by clicking on the navigation button for the *Student Results* page. They could not view other assessors' marking results to avoid bias.

WACE Practical score

The Western Australian Certificate of Education is the qualification that needs to be obtained by Year 12 students in Western Australia upon completion of their secondary school studies. WACE examination was regulated by the Curriculum Council of Western Australia at the time when this study commenced. To obtain the WACE, Year 12 students must satisfy the WACE requirements. For both Design and Visual Arts courses, the WACE examination in 2011 consisted of two components: theoretical and practical. The present study was only concerned with the practical score.

For WACE practical examination for Design the students submitted a Design portfolio and a resolved artwork for the Visual Arts course. WACE assessors marked these student works in a double-blind marking process. In the marking process each assessor scored anonymous student work independently. A reconciliation meeting in each course was arranged to discuss student works which were given scores with a large difference. In this reconciliation process the assessors examined the works together to agree on a score.

Assessor interview

After the scoring processes were concluded, an interview was conducted with each assessor from both the Analytical marking and the Comparative Pairs judgements. In this interview the assessors were questioned about the authenticity and quality of the digital representations, the scoring tool, and the quality of student work. This interview was a structured interview using interview questions that were adapted from previous CSaLT research projects that have been tested for validity. Several interviews were conducted in person and several others were conducted through emails because of time and distance limitations. The interview questions are presented in Appendices E and F.

Assessor notes from the ACJ system

As was described in *Figure 3.8*, the ACJ system included notes boxes for assessors to record their judgement notes. The assessors could make notes on individual portfolio and each pair of portfolios. Notes on individual portfolios could refer to the quality of the portfolio, for example the techniques or the material selected in Visual Arts and innovation or design solution in Design. Notes on the pair could refer to the comparison made on the pair of portfolios presented, for example the components that made portfolio A to be the winner. Besides its use to assist the assessors to remember their judgements, these notes also provided data on judgements. These notes specified the components and quality that were considered important and were used to be the deciding factor when two portfolios were of similar quality. In this study, these assessor notes were used in the discrepancy analysis that is discussed later in this chapter.

Data analysis framework

Data analysis framework for this study was focussed on the research questions of the study, as is the nature of pragmatic research. Data were obtained from the three scoring processes and from interview with the assessors from the main project. *Figure 3.11* depicts the data analysis framework used in this study.

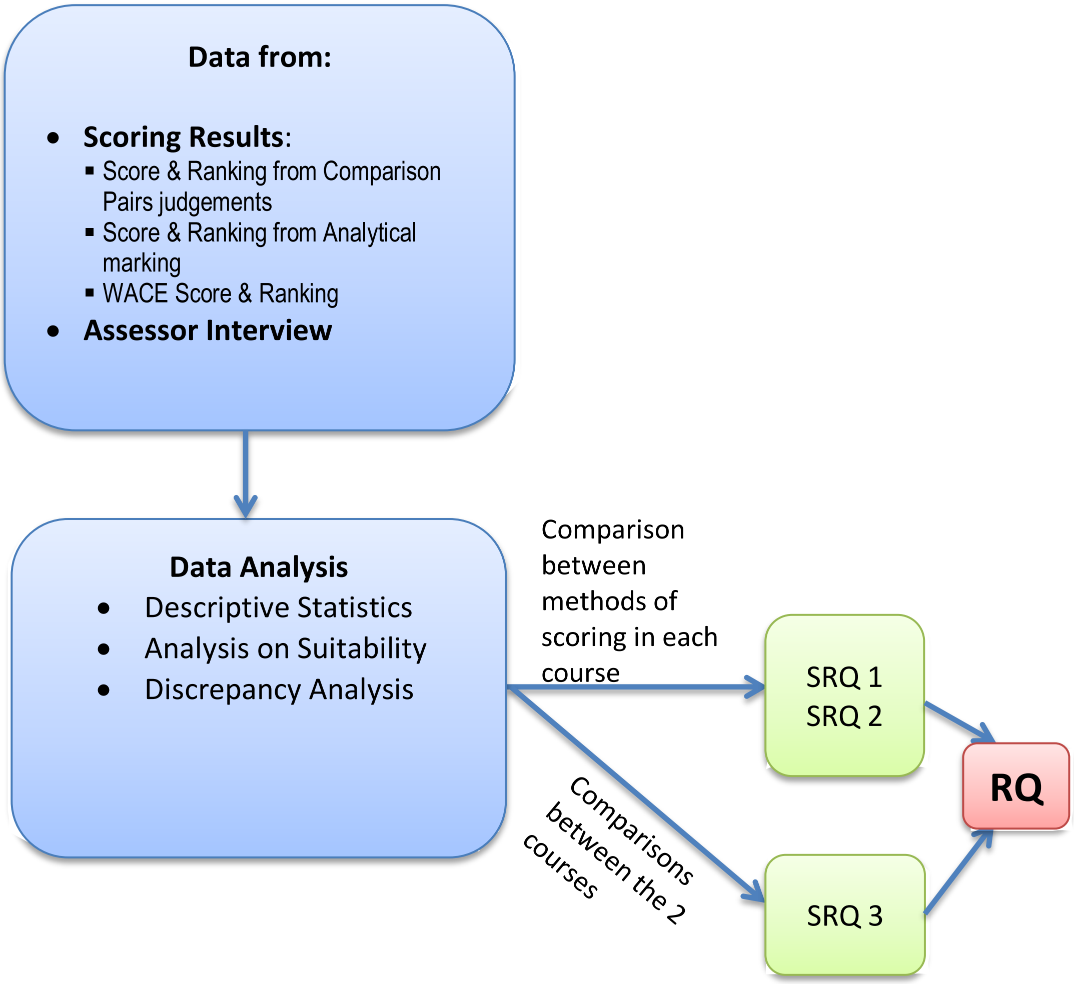

Data from:
Scoring Results:
Score & Ranking from Comparison Pairs judgements
Score & Ranking from Analytical marking
WACE Score & Ranking
Assessor Interview
Data Analysis
Descriptive Statistics
Analysis on Suitability
Discrepancy Analysis
Comparison between methods of scoring in each course
Comparisons between the 2 courses
SRQ 1
SRQ 2
SRQ 3
RQ

Figure 3.11 Data analysis framework.

Quantitative data from the scoring methods were managed and processed using excel spreadsheet and SPSS software to generate data for further analysis. The descriptive statistics regarding the scores and rankings obtained from the three scoring methods were presented to provide information on the results from this preliminary data analysis.

The analysis on suitability was based on three points of reference, which were:

- reliability
- comparison with other scoring methods
- validity issues.

The reliability of the results from the Comparative Pairs Judgements method was obtained through the ACJ system in the form of reliability coefficients of the Rasch model that was similar to the Cronbach's alpha. Unlike the Cronbach's alpha that only represented the internal reliability of the results, however, these coefficients also represented the inter-rater reliability.

The reliability of the results from the Analytical marking method was obtained through a calculation on item reliability using the Cronbach's alpha on SPSS, a statistics software. This reliability represented the consistency of the marking results based on the criteria. The inter-rater reliability for this marking method was calculated by using correlations between assessors. This approach for calculating inter-rater reliability was called the *consistency estimate* (Stemler, 2004), which was a reliability estimate that only considered the consistency between assessors without taking into account the mean or median values of each assessor. As such, a high consistency estimate of inter-rater reliability does not mean the assessors assigned similar scores to the students, only that the assessors agreed on the ranks of the students.

As was discussed in Chapter 2, the validity analysis for this study was based on the validation framework developed by Kane (2006) and Shaw et al. (2012), and by using Pollitt's inferences on threats to validity. Only one inference from the framework, *scoring*, was considered relevant to this study and therefore was used as a guideline to analyse the quantitative and qualitative data. Evidence for validity as well as threats to validity derived from data analysis were investigated and discussed.

Results from the validity analysis, together with the descriptive statistics, provided information on the quality of the Comparative Pairs judgements in each course. This analysis was related to the first subsidiary research questions (SRQ 1). A comparison analysis between the two online scoring methods provided information on the second subsidiary research questions (SRQ 2). Subsequently, a comparison analysis between the two courses provided information on the third subsidiary research questions (SRQ 3). Results from all above processes provided information on the overarching research question. The general analysis of data is presented in chapter 4 for Design and in chapter 5 for Visual Arts. The cross-case analysis between the results for the two courses is presented in chapter 6.

Ethical considerations

In principle, the general purpose of the ethical consideration of scientific research could be connected to "protecting individual autonomy" (Howe, 1999, p. 22). As Howe further explicated, the deontological framework of ethics does not justify objectifying people for the sake of research. Over the years, this ethical framework gradually developed into more comprehensive principles aimed to safeguard individual autonomy.

Based on this deontological view of research ethics along with other ethical views such as consequential view, virtue, situational view, research institutions and governments around the world constructed guidelines to regulate research ethics (Cohen et al., 2011; Israel & Hay, 2006; Shrader-Frechette, 1994). These guidelines encompass potential ethical issues such as informed consent, research procedures, data access and confidentiality, anonymity, cost and benefit, conflicts of interest, bias, sensitive social and political data and many more.

As in any other scientific research, ethical issues could be present in this study. In order to limit these issues, this study took measures to comply with ethical guidelines from the university. Ethics clearance was lodged and obtained through the ECU Human Research

Ethics Committee (HREC). The application contained an overview of the study, information letters for the participants including a consent form and data collection instruments. Furthermore, because this study was conducted within the main project, permission request to access the data collected in the project was also included in the application. Correspondingly, for the main project, the ethics clearance was applied from Edith Cowan University, the Department of Education Western Australia, and the Catholic Education Office of Western Australia because schools from all three sectors were involved in this project. For independent schools the Principal determined whether the school participated in the research project.

Anonymity and confidentiality

As a measure to maintain the privacy of the participants in this study, every school was given an identification code that only the researchers of the project were informed. Each student who agreed to be involved was also given an identification code that was connected to their school after they returned the parent and student consent forms that they and their parent signed.

All data that were collected were either kept in a locked filing cabinet in CSaLT office or saved on CSaLT server accessible only by the researchers through their paassword. These data would be destroyed or deleted only after seven years after the final report for the project was sent to the stakeholders.

Summary

This chapter discussed the research method employed in this study. The participants, the nature of data collected, and the data analysis process were described. The following two chapters, Chapter 4 and Chapter 5, present the analysis of data for the Design and Visual Arts course samples consecutively.

CHAPTER 4
FINDINGS FROM THE ANALYSIS OF DATA - DESIGN

This chapter presents the results from an analysis of the data collected for the Design course assessment. This starts with a description of the portfolios submitted, followed by a presentation of the results of an analysis of the data from the Comparative Pairs judgements method and the Analytical marking of the digitised copies of the portfolios, as well as a comparison with the data from the official WACE analytical marking. Then a discrepancy analysis between the scores from the different sources is discussed. Next a description of the assessor interview data is presented, combined with an analysis of the reliability of the scores, and validity of the Comparative Pairs judgements. The purpose of this presentation of results is to provide findings with which to address the research questions in a later chapter.

Student Work

The practical task for the Design course WACE examination was submitted as a 15-page A3 paper portfolio for each student to the Curriculum Council for marking. The portfolios belonging to the students who participated in the main project were then scanned and saved as a PDF file digital portfolio. The scoring data that were analysed in this study were the results of the scoring processes of the digital and original paper portfolios. The original portfolios were used in the WACE practical examination process along with the rest of the Design portfolios of the entire cohort while the digital portfolios were used in both the Analytical marking and Comparative Pairs judgements processes. In this section, factors that might affect the scoring data from the nature of student work, the digitisation processes, and the technical limitations are discussed.

Nature of student work

For the Design course the practical component of WACE was in the form of a 15-page A3 paper portfolio that consisted of different components, as was described in Chapter 3.

These portfolios were meant to show “evidence of the design process used to arrive at completed design solutions” (Curriculum Council of Western Australia, 2010a, p. 35). Thus for the Design course, the type of the assessment task was evidence of process, in contrast to the assessment task in the Visual Arts course, which was a finished product. This portfolio could be printed, hand written, or hand drawn on 15 pages of A3 paper. The researchers for the overarching main project scanned these paper portfolios, and the resulting PDF copy was saved in a digital folder, for the assessors to view, in both the Comparative Pairs judgement and the Analytical marking processes.

In their 15-page Design portfolios, the students had to present the complete stages in the development of up to three design projects in the form of mind maps, sketches, descriptions, photographs, and others. These requirements could make the portfolios difficult to assess. The nature of the portfolio presented challenges in the scoring processes. In the Analytical marking process the assessors had to find the evidence for each criterion in the analytical marking rubric from the 15 pages of varied types of work then assign a score. Finding details that represented the evidence for each criterion all throughout the portfolio could make the scoring process complicated, and in turn compromise the reliability of the scoring results. In the Comparative Pairs judgements process the assessors compared two portfolios based on the holistic criterion. While in this process the assessors did not have to match different criteria to a range of evidence that could be contradictory in some cases, deciding which portfolio was a better one was still not a straightforward task, especially when the quality of the two paired portfolios was similar or when one portfolio was stronger in one component but weaker in another. In cases like this, there was a possibility that the visual aspect of the portfolio or the assessors’ personal preference tipped the scale, and the reliability of the results could be affected.

Constraints from the digitisation process

During the digitisation process of the Design portfolios, there were several issues that could possibly compromise the quality of the scoring and perceptions data. The main digitising issues were time constraints and the variety of materials used by the students. This section discusses these issues.

Time constraints

The scanning process of the student work was limited by time constraint. The whole digitisation process needed to be finished within the two-day window period between the submission day and the official marking day. The Curriculum Council marking process was scheduled to start two days after the submission day and once it started the student work was not available to the researchers. Because of the time available to digitise the portfolios was quite limited, when the researchers encountered problems with the digitisation process such as when the pdf copy was not clear or appeared in different colours to the original one there was not enough time to find a solution. The disparity between the original portfolios and the digital version could affect the scoring results. For the Design course, this constraint was not too significant compared with Visual Arts because the scanning process was simple and portfolios were typically in school bundles so it was easy to find the work of the students participating in the study. All Design portfolios were successfully scanned within the time period.

Types of materials used

Beside time constraints, problems also arose from the variety of materials used by the students for their portfolio. The variety of the submitted work included copy paper, card paper with varied thickness, thin tracing paper, and glossy paper. While the appearance quality of most of the digital portfolios was quite similar to that of the submitted work, there was a notable difference in some portfolios, especially in the ones printed on glossy material. The scanning of this type of material could either create smudges or differences in colour in the digital version. Aside from that, the automatic feeder of the scanner was not designed for materials that were too thin or too thick; therefore these types of portfolios were more time-consuming because they needed to be scanned individually. These varied types of materials, combined with the limited time available, could affect the clarity of the digital portfolios as well as create disparity between the original portfolios and the digital version, which in turn could affect the scoring results.

Technical limitations

Even though two commercial quality digital scanners were used to digitise the Design portfolios, several portfolios, when scanned, appeared to have different colours to the original. Beside the problem with the scanned portfolios, there was also a problem with the file size. Because the portfolios consisted of 15 A3-size pages and the highest quality scans were employed, the digital file size of the scanned portfolios was mostly quite large. Therefore, when the portfolios were opened for judgements, on a few occasions it took several minutes for the files to open, especially when the assessor’s Internet connection was slow. This problem could result in longer and more difficult judgement processes.

Analysis of the Scoring Data

This section presents the results from an analysis on the scoring data from the three sources, which were the Analytical marking of the digital portfolios, the official WACE analytical marking, and the Comparative Pairs judgements. Several Design course educators from secondary and tertiary levels were invited to be assessors for both the Analytical marking and the Comparative Pairs judging and ten decided to be involved. Two of these assessors, who were experienced WACE markers, were also the two assessors in the Analytical marking process.

Analytical marking for Design course

For the Design course two assessors marked the digital portfolios analytically by using the rubric in the online tool that was also used to obtain the WACE practical examination score from the paper portfolios (marked by others). The rubric consisted of six criteria with maximum score points ranged from 6 to 10 for each criterion with a total score of 50. The complete rubric that was used in these two scoring processes can be viewed in Appendix C. The criteria were:

C1: Design elements and principles (0 – 6)
C2: Design Process (0 – 6)
C3: Analysis and Innovation (0 – 10)

C4: Experimentation and Selectivity (0 – 10)

C5: Production knowledge and skills (0 – 10)

C6: Communication and visual literacies (0 – 8)

Processes and time taken for marking

The Filemaker Pro scoring interface was equipped with a timer to record the time spent by the assessors to assess the portfolios. This timer started at the beginning of each scoring session and recorded the total time each assessor needed to assess a portfolio. The recorded time could include unintended breaks during the scoring as well as the time needed for the portfolios to load. The size of the digital files was quite big and therefore it could take some time to load. In average, the assessors spent 6.4 minutes to mark a portfolio. The total amount of time for marking the 82 Design portfolios by the two assessors was 17.5 hours. The shortest time recorded was five minutes and the longest was 15 minutes per portfolio. Considering the file size of the portfolios was relatively similar, the portfolios that needed more time could be the ones that were difficult to score. This difficulty could be caused by the lack of clarity of the pdf version, contradicting qualities of the portfolios within the criteria, or missing portfolio components.

Scores from marking

The Filemaker Pro scoring interface provided the marking rubric for Design on which the assessors assigned a score to each criterion for each portfolio. Results from this marking process were recorded in the Filemaker Pro database then imported to a spreadsheet and analysed by using SPSS. For each student a score for each criterion was recorded and then summed to generate a total score. The structure of these scores from the Analytical marking is as shown in Table 4.1. This structure was designed to make analysing the scoring results based on criteria, assessors, or schools easier.

Each school participating in the study was assigned an identification code that consisted of two letters, the first letter was D, for Design, the second letter was the school code. Each student participating in the study was assigned an identification code that consisted of two letters from the school code followed by three digits of number. The purpose of this coding was to maintain the privacy of both the schools and the students involved. Only the researchers involved in the project had access to this coding. There were 82 students from six schools involved in this project, with the number of students varying from only four in school DB to as many as 21 in school DT.

Table 4.1
Structure of Analytical Marking Data

ID	C1 (6)	C2 (6)	C3 (10)	C4 (10)	C5 (10)	C6 (8)	Total (50)
DB 903	5.0	4.5	7.5	7.0	9.0	6.5	39.5
DB 905	4.5	5.0	8.5	8.5	9.0	7.0	42.5
DB 906	4.0	4.5	6.5	6.5	7.5	6.0	35.0

Analysis of scores based on schools

Table 4.2 presents the analysis of scores based on schools. This analysis was intended to examine possible patterns or peculiarity among schools in each criterion. In this analysis, the means and standard deviations for each school in each criterion were calculated and compared. In general there was no school with particularly different mean scores or standard deviations across criteria. All schools had mean scores within two standard deviations difference with the average mean score of each criterion.

Table 4.2
The Mean Score for Each School per Criterion from the Analytical Marking Process

School	N	Score (SD)						
		C1 (6)	C2 (6)	C3 (10)	C4 (10)	C5 (10)	C6 (8)	Total (40)
DB	4	4.3 (0.6)	4.6 (0.3)	7.4 (0.9)	7.4 (0.8)	8.1 (1.0)	6.3 (0.6)	38.0 (3.6)
DL	18	3.3 (0.7)	3.4 (0.6)	4.8 (1.2)	4.8 (1.1)	5.4 (1.1)	4.1 (1.0)	25.9 (5.3)
DM	13	4.1 (0.7)	4.0 (0.9)	6.2 (1.3)	6.3 (1.2)	6.1 (1.2)	5.4 (1.0)	32.2 (5.8)
DN	17	4.4 (0.7)	4.5 (0.6)	7.0 (0.8)	7.0 (0.8)	7.0 (1.0)	5.8 (0.8)	35.7 (4.2)
DT	21	3.6 (0.6)	3.9 (0.6)	5.5 (1.2)	5.9 (1.1)	5.5 (1.0)	4.4 (0.9)	28.8 (5.0)
DV	9	3.4 (0.8)	3.6 (0.7)	4.9 (1.4)	5.1 (1.1)	5.3 (1.0)	4.3 (0.8)	26.6 (5.4)
MEAN	82	3.8 (0.8)	3.9 (0.8)	5.8 (1.4)	5.9 (1.3)	6.0 (1.3)	4.9 (1.2)	30.3 (6.3)

Table 4.3 shows the same scores in percentage. This conversion helps in making differences among schools or criteria more pronounced. As is shown in Table 4.3, the total mean scores in each criterion ranged between 58.0% and 65.5%, with criterion C3 -

Analysis and Innovation being the lowest and C2 - *Design Process* the highest. Overall the mean in each criterion was around the total mean of 60.7% (SD=12.6%) with no criterion having more than one standard deviation difference to the mean. This means that while there was variation in the way each criterion contributed to the total school score, there was no particular criterion that contributed too little or too much to the total school score.

Table 4.3
The Mean Score in Percentage for Each School per Criterion from the Analytical Marking Process

School	N	Score (%)					
		C1	C2	C3	C4	C5	C6
DB	4	70.83	77.08	73.75	73.75	81.25	78.13
DL	18	55.09	56.94	48.33	48.06	54.44	51.74
DM	13	67.95	66.67	62.31	63.08	61.15	67.79
DN	17	73.04	75.00	70.00	69.71	70.29	72.79
DT	21	59.52	64.68	55.00	58.57	55.48	55.36
DV	9	57.41	60.19	48.89	51.11	52.78	53.47
MEAN	82	63.0 (13.0)	65.5 (12.8)	58.0 (14.1)	59.2 (13.1)	60.2 (12.8)	61.1 (14.5)

The mean score percentage for each school was quite consistent across the six criteria with only a few slight exceptions. This could be observed in *Figure 4.1*. The schools' means in each criterion formed a distinctive pattern. Schools DB, DN and DM were above the overall means in all criteria with school DB had the highest mean scores in all criteria except in C1 - *Design Elements and Principles*. Meanwhile, schools DT, DV and DL were below the overall means in all criteria. This pattern could indicate agreement among criteria, which contributes to the construct validity of the assessment. However, further analyses needed to substantiate this claim were not conducted because it was not relevant to the aim of this study, which was focussed on the Comparative Pairs judgements.

This pattern could also indicate the influence of school culture; such as collective academic characteristics of Design students in each school (e.g. persistence, understanding, intelligence), specific teaching methods, teaching-to-the-test approach, availability of school facilities, and others; on student achievement in practical assessment that could be an interesting and important topic for a further study. It should be noted, however that the sample size is quite small for each school, for example school DB only had four students participating in the study.

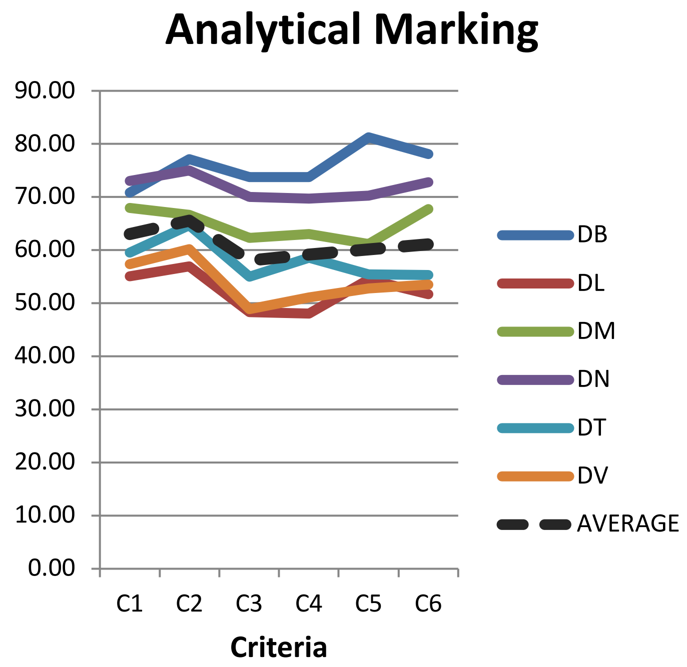


Figure 4.1 Analytical marking result for each school per criterion.

Analysis of scores based on assessors

A summary of the scores obtained from each assessor in each criterion is presented in Table 4.4. In general there was reasonable agreement between the two assessors in terms of score range, however, Assessor 2 tended to utilise a wider range of scores than Assessor 1 while the mean scores given by Assessor 1 were slightly higher than Assessor 2 in several criteria. The standard deviations of scores given by the two assessors were relatively similar, indicating that the spread of the scores given by both assessors in each criterion was quite regular.

Table 4.4

Descriptive Statistics on Marking for All Students by Each Assessor

Set of Criteria	Assessor	Possible	Range	Mean	Std. Deviation	Mean (%)*
C1 (0-6) Design elements and principles	1	6	2-6	3.7	1.0	62.0
	2	6	2-6	3.8	0.9	64.0
	Average	6	2-5.5	3.8	0.8	63.0
C2 (0-6) Design Process	1	6	2-6	3.9	1.0	65.4
	2	6	0-6	3.9	0.9	65.7
	Average	6	1.5-6.0	3.9	0.8	65.5
C3 (0-10) Analysis and Innovation	1	10	2-9	6.0	1.8	59.9
	2	10	2-9	5.6	1.5	56.2
	Average	10	3-8.5	5.8	1.4	58.0
C4 (0-10) Experimentation and selectivity	1	10	2-9	6.2	1.7	61.5
	2	10	2-9	5.7	1.4	57.0
	Average	10	3-9	5.9	1.3	59.2
C5 (0-10) Production knowledge and skills	1	10	2-9	6.2	1.5	61.6
	2	10	2-10	5.9	1.5	58.8
	Average	10	2.5-9.5	6.0	1.3	60.2
C6 (0-8) Communication and visual literacies	1	8	2-7	5.0	1.4	62.5
	2	8	2-8	4.8	1.3	59.6
	Average	8	2-7	4.9	1.2	61.1

*Percentage of the mean average

When the scores given by the two assessors were compared, the largest score difference was 26 (out of 50) with a mean of 5.6 (SD=4.4). The correlation between the two assessors' scores, as well as between individual assessor scores and the WACE scores was significant but relatively low. The correlation coefficient for the scores given by the two assessors was 0.53 ($p<0.01$), indicating that even though both assessors were experienced Design assessors and were using the same Analytical marking criteria, the agreement between them was only moderate, which consequently indicated moderate inter-rater reliability for the Analytical marking. *Figure 4.2* shows the scatter plot of the scores given by the Analytical assessors.

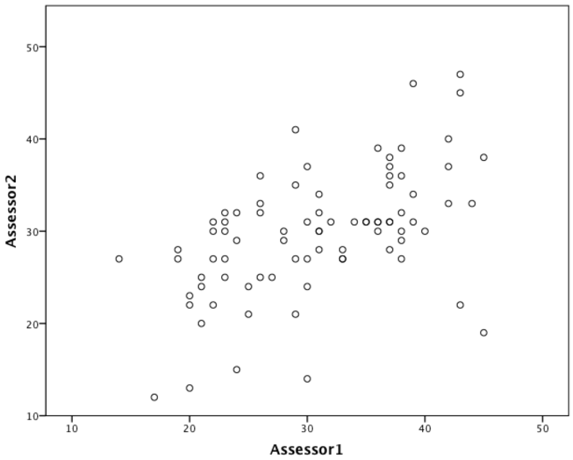


Figure 4.2 Scatter plot between scoring results from the Analytical marking assessors.

Even though the correlation between scoring results from the two assessors were low, there were not many portfolios with large differences between the scores. There were only three out of 82 portfolios (3.7%) with a difference of more than 2 standard deviations to the mean, which were DV901, DV904 and DV906. These portfolios were all from one school, DV. When the scores were ranked, however, the differences became much larger. This is discussed later in this chapter.

Comparison between Analytical and WACE practical Marking scores

The WACE practical scores for participants in the study were provided by the curriculum authority. These scores were generated from assessors who marked the students' paper portfolios using the same rubric to that used in the study for the Analytical marking of the digitised portfolios. That is, the difference between these two scoring methods was only the form of the portfolios being marked. In the official WACE marking the assessors marked the original printed portfolio while in the Analytical marking for the study, the assessors marked the digitised version.

As for the Analytical marking, in WACE marking there were several assessors with each portfolio being marked by at least two assessors. In case of extreme dissimilarities in marking, a meeting was held to discuss the differences and to obtain an agreed score. The WACE practical score used in this study was the mean of the scores from the assessors, or the score from the reconciliation meeting. A summary of the results from the Analytical marking and WACE marking is shown in Table 4.5.

Table 4.5
Descriptive Statistics on Analytical Marking and WACE

		N	Range	Mean	SD
Analytical Marking	Assessor1	82	14.0 – 45.0	30.9	7.7
	Assessor2	82	12.0 – 47.0	29.7	6.8
	Average	82	14.5 – 45.0	30.3	6.3
	Average (%)	82	29.0 – 90.0	60.7	1.6
WACE	WACE Practical	82	15.0 – 50.0	35.2	8.2

Compared to the result from the WACE practical marking, the mean from the Analytical marking was considerably lower. In general, the WACE markers utilised a wider range of scores, as the Analytical marking maximum score was only 45.0 while that of WACE was 50.0 with the minimum scores quite similar. Correspondingly, the score distribution from the WACE practical marking was also relatively more widely spread than the Analytical marking with a standard deviation of 8.2, which was considerably higher than the standard deviations of the scores from both Analytical marking assessors which were 7.7 and 6.8 consecutively.

Comparative Pairs judgements for Design course

Data from the Comparative Pairs judging were obtained from the ACJ system using the judgements done by ten Design assessors. All assessors were either qualified and experienced teachers in the Design course or academics in Design. Three of them were involved in the WACE examination marking. More information on the assessors is discussed in the Assessor Interview Data section.

Before the scoring process started the researchers involved in the project hosted a four-hour workshop with the Design assessors. This workshop had two main purposes; the first was to decide on a holistic criterion upon which the Comparative Pairs judging was to be based. This criterion was based on the marking rubric developed for the official WACE practical examination. In this workshop, the assessors discussed and decided on an holistic criterion for the Comparative Pairs judgement which was:

Holistic Criterion: Judgement about performance addresses students' ability to apply elements and principles of design in recognising, analysing and solving specified design problems innovatively with consideration for a target audience and justify design decisions through experimentation and production.

The WACE marking criteria upon which this holistic criterion was based were:

- *Design elements and principles* - Application of design principles, use of design elements
- *Design process* - Brainstorming, idea generation methods, visualisation of concepts
- *Analysis and innovation* - Analysis of information and translation into design concepts, application of interrelated thinking and innovative development process
- *Experimentation and selectivity* - Use of interpretive skills and problem solving
- *Production knowledge and skills* - Selection and use of a diverse range of skills, techniques and procedures, application of planning and production methods
- *Communication and visual literacies* - Ability to interpret design brief, ability to construct a visual image that conveys a message

These criteria were also the criteria that were used in the Analytical marking process for the digital portfolios.

The second purpose of the workshop was to introduce to the assessors the judging interface of the ACJ online system and to ensure that there was a common understanding on how to use the holistic criterion. At the end of the workshop the assessors started judging the first few pairs in the first judging round. The rest of the judging process was conducted off-site at home or workplace.

ACJ System Data on Comparative Pairs Judging

The Comparative Pairs judgements data were obtained from the ACJ system. The system created the pairings from which the assessors judge the better one in each and subsequently ranked the students based on those judgements. At the end of the whole judgement process, which consisted of several rounds, the system ranked the portfolios on a parameter measurement scale in Rasch logits, and provided information on judgements sessions as well as an analysis of reliability and individual portfolio or assessor misfits. Features from the ACJ system that were used in this study were discussed more fully in Chapter 3.

In the first rounds, the ACJ system paired the portfolios randomly then more adaptively, resulting in gradually faster judgements and more accurate scoring results. *Figure 4.3* shows how the standard error bars of the parameter values for the portfolios improved between the first and the last round. The graph curve also became smoother, which indicated that the rank of the student was getting more closely together and the difference in quality became finer.

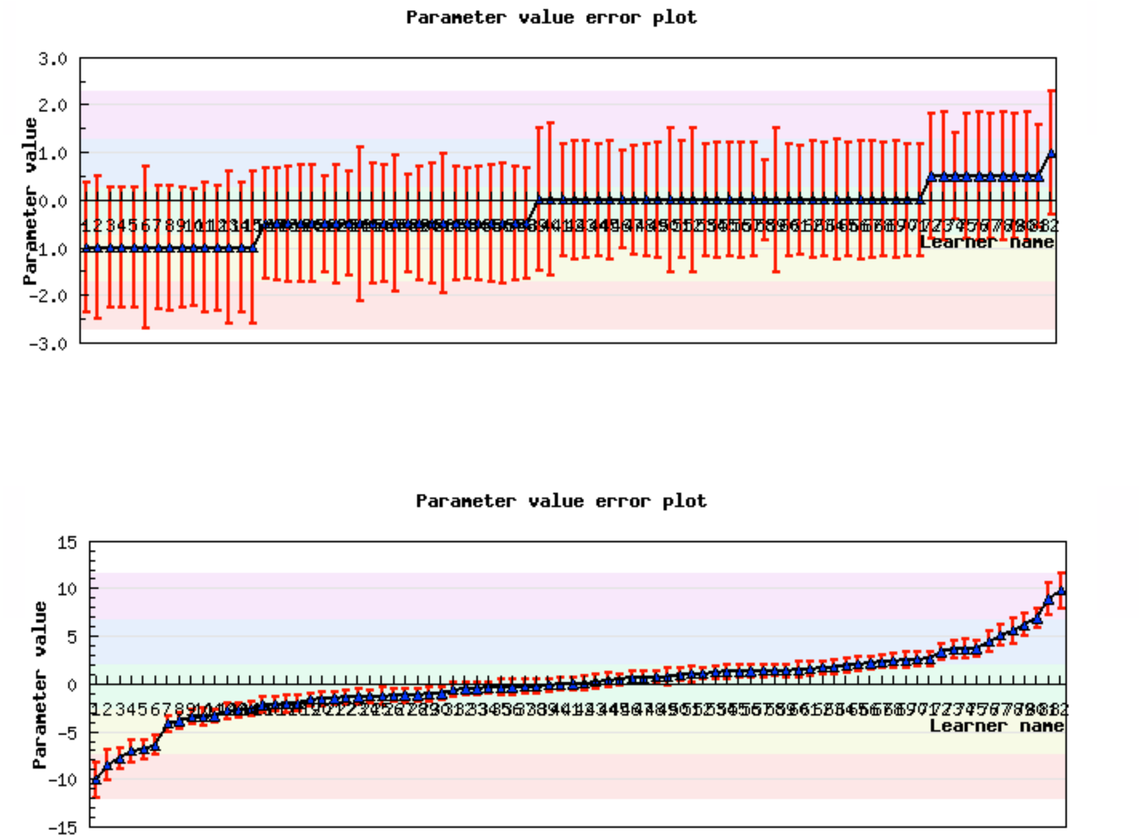


Figure 4.3 Parameter value error plot from the first and last rounds.

By the end of the thirteenth round, the reliability coefficient reached 0.941, and the judgement process was concluded because it was understood that after this point it was likely that there would be little increase in the reliability coefficient. This high reliability level represented both the internal reliability and the inter-rater reliability of judgement among judges (Kimbell, 2007).

Table 4.6 shows how the reliability coefficient increased for every round of judgement. Related to the discussion of the ACJ system in Chapter 3, the first six rounds had not resulted in a meaningful reliability coefficient, therefore it is not included in the table. As can be seen in Table 4.6 below, there was a jump in the reliability coefficient between the sixth and the seventh rounds. From the seventh round onward there was a relatively steady increase in the reliability coefficient as more fine-tuning in the pairing was created and portfolios of more similar quality were paired to be judged.

Table 4.6

Reliability Coefficients from the Last Eight Rounds of Comparative Pairs Judgements

Round	r
6	0.610
7	0.836
8	0.867
9	0.894
10	0.910
11	0.926
12	0.936
13	0.941

Consistency of the Assessors and Judgements

During the judgement process, the ACJ system compared each judgement made by the assessors with the overall judgements. This process provided the researchers with information on the consistency of the assessors in misfit statistics data. These misfit data included the mean residual, the weighted mean square, and the unweighted mean square. The consistency statistics from the ACJS is as shown on Table 4.7.

The mean residual for each assessor, except Assessor 10 who only did 17 judgements, was around the mean of 0.44. This shows that in general the assessors' judgements were consistent with one another. The misfit statistic shown by the weighted mean square had a mean of 1.21 (SD=0.12) with only two assessors (Assessors 5 and 7) had a mean difference that was slightly more than one standard deviation from the mean. This further indicated that there was no extreme inconsistency between assessors in the judgements. Among all 543 judgements there were only 25 (4.6%) judgements that the system identified to be inconsistent.

Table 4.7

Consistency Statistics for Assessors for the Design Portfolios

Assessor	Count	Mean Residual	Unweighted mean square	Unweighted Z	Weighted mean square	Weighted Z
1	1	0.50	1.00	0.00	1.00	0.00
2	69	0.42	4.41	1.86	1.20	2.25
3	42	0.46	1.11	0.42	1.16	1.55
4	69	0.46	6.49	3.35	1.32	3.52
5	69	0.48	13.38	3.79	1.39	5.06
6	69	0.45	0.97	8.93	1.02	0.42
7	69	0.45	14.60	2.80	1.35	3.80
8	69	0.45	1.69	2.76	1.26	2.80
9	69	0.45	1.16	1.80	1.22	2.50
10	17	0.33	1.06	1.14	1.18	0.73
	Mean:	0.44	4.59	2.68	1.21	2.26
	S.D.:	0.04	5.02	2.39	0.12	1.53

Processes and Time Taken for Judging

Ten judges were involved in the Comparative Pairs judging, however, there was not enough activity from one of them (Assessor 1), therefore only results from the other nine assessors were analysed. There were 543 judgements in 50 hours made in total, averaging at 5:36 minutes per judgement. Each judgement took from 2.53 to 11.21 minutes per judgement, with fluctuating average time. It should be noted that this amount of time could include breaks that might be taken by the assessors during judgement sessions. However, the system calculations tried to make allowances for extreme values. Table 4.8 shows the estimated time for each round in the Comparative Pairs judgement process.

Table 4.8

Estimates of Time Taken Making Judgements for Comparative Pairs Judging of the Design Portfolios

Round	Total time (hrs)	Judgements	Average Time per Judgement (hrs)
1	3:54:19	40	0:05:51
2	2:50:44	31	0:05:30
3	4:12:32	36	0:07:00
4	3:00:09	39	0:04:37
5	3:33:45	39	0:05:28
6	4:39:55	39	0:07:10
7	2:40:00	39	0:04:06
8	1:50:40	36	0:03:04
9	2:57:43	39	0:04:33
10	2:29:50	39	0:03:50
11	3:24:01	40	0:05:06
12	2:05:16	39	0:03:12
13	1:17:58	41	0:01:54

Scores from Comparative Pairs Judgements

Scores from the Comparative Pairs judgements were obtained from the ACJ system. At the end of the judgement session the ACJ system provided a summary of the final location parameter for each student, including the inconsistency statistics. The structure of this summary was as displayed in Table 4.9.

Table 4.9

Sample of Student Location Parameter Result from the ACJ System

Student ID	Parameter	SE	Unweighted mean square	Unweighted Z	Weighted mean square	Weighted Z
DB903	1.03251	0.82	1.21	1.18	1.16	1.08
DB905	8.88936	1.63	0.48	46.91	0.85	-0.69
DB906	1.38184	0.68	30.44	4.06	2.03	2.98
DB909	2.15786	0.70	0.82	1.25	1.02	0.17
DL901	-1.63479	0.66	1.37	0.74	1.31	2.10
DL903	-3.47411	0.91	17.98	17.97	1.96	3.23
DL907	-2.07281	0.81	1.83	0.96	1.73	2.96
Mean:	4E-07	0.79	5.43	3.56	1.24	1.30
S.D.:	3.43333	0.27	11.68	7.71	0.28	1.26

This judgement and analysis process resulted in a score set that ranged from -10.085 to 3.454 logits. The frequency distribution of the location parameter had a mean of

0.0000004, which was very close to 0 as expected in a normal distribution. The graph of the frequency distribution is displayed in *Figure 4.4*. This location parameter was based on the Rasch dichotomous model that was employed by the ACJ system as discussed in Chapter 3. From the 82 portfolios assessed in this scoring method, six portfolios (7%) had a weighted mean square value above 2 SD from the average value. This suggested that the judgements were less conclusive on these six portfolios than the rest.

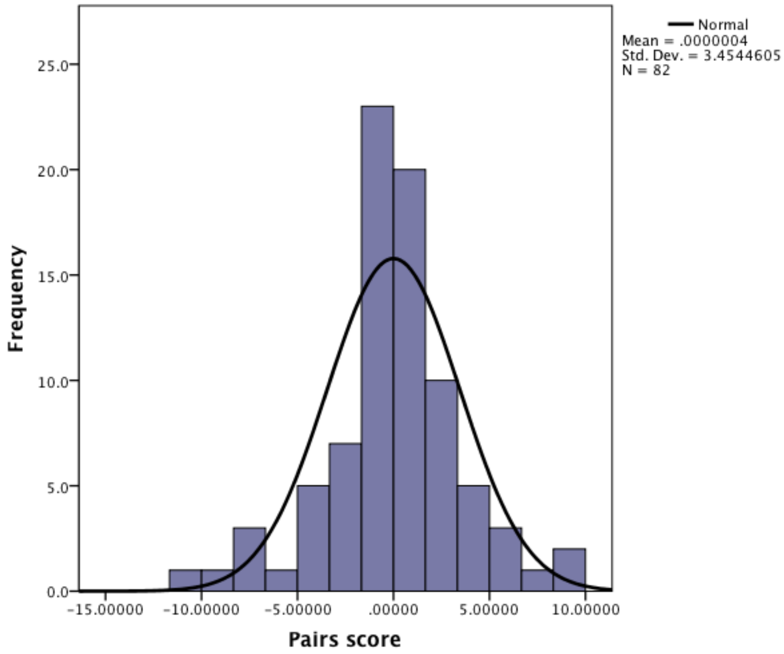


Figure 4.4 Frequency distribution of CP scores.

The ACJ system judged that, assuming the scores represented a population of about 6 SD's wide and that bands 3 SE's apart are distinguishable, there were up to 8.6 reliably distinct bands. These bands could be used for grading the portfolios but because they were not pertinent in this study, they are not discussed. A further normality test, however, showed that the parameter distribution did not follow an exact normal distribution even though it was not skewed, as indicated in Table 4.10.

Table 4.10

Normality Tests Results

Descriptives

			Statistic	Std. Error
Pairs score	Mean		0.0000004	0.38148135
	95% Confidence Interval for Mean	Lower Bound	-0.7590279	
		Upper Bound	0.7590286	
	5% Trimmed Mean		0.0270798	
	Median		0.0181436	
	Variance		11.933	
	Std. Deviation		3.45446050	
	Minimum		-10.08510	
	Maximum		9.76796	
	Range		19.85306	
	Interquartile Range		3.23737	
	Skewness		-0.177	0.266
	Kurtosis		1.485	0.526

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Pairs score	0.099	82	0.048	0.963	82	0.020

a. Lilliefors Significance Correction

The z-score for skewness for this parameter is:

$$z = \frac{Skewness}{Std.Error} = \frac{-0.177}{0.266} = -0.665$$

This skewness z-score is within the range of ± 1.96 (p<0.05), indicating that the distribution could be considered symmetrical and not significantly skewed. The z-score for kurtosis, on the other hand, is:

$$z = \frac{Kurtosis}{Std.Error} = \frac{1.485}{0.526} = 2.823$$

This kurtosis z-score is outside of the range of ± 1.96 (p<0.05), indicating that the distribution was leptokurtic with most portfolios were located around the average value as could also be seen in the histogram in Figure 4.5. The significant probability value of the Shapiro-Wilk statistic is 0.02, which is lower than α= 0.05, suggesting that the parameter distribution is not normally distributed.

Summary of scoring results

Shown in Table 4.11 is a summary of the mean scores and ranks from all three methods of scoring for each school. The location parameter (logits) resulted from the Comparative Pairs judgement was not in the same scale as the other two scoring methods, therefore the result was rescaled using the mean and standard deviation from the Analytical marking result. The Analytical marking result was used as a baseline because the Comparative Pairs judgements method used a holistic criterion that was based on the criteria for the Analytical marking, and also because unlike the WACE marking process, both the Analytical marking and the Comparative Pairs judgements processes used the same digital portfolios. The Comparative Pairs judgements scores were in logits as the ACJ system used the Rasch dichotomous model to calculate the student ability and criterion difficulty. These scores were rescaled and ranked to parallel the results from the other scoring processes to better describe and compare the results.

Table 4.11
Scoring Result Summary from All Scoring Methods for Each School

Case School	N	Analytical		CP		WACE Practical	
		Score (50)	Rank	Logits (Rescaled)	Rank	Score (50)	Rank
DB	4	38.0 (3.7)	11.5 (8.2)	36.5 (6.8)	18.8 (12.7)	39.5 (5.1)	27.9 (17.8)
DL	18	25.9 (5.3)	58.4 (18.8)	22.8 (5.0)	71.3 (8.2)	28.2 (6.1)	61.2 (14.3)
DM	13	32.2 (5.8)	35.1 (22.6)	28.5 (4.0)	50.6 (16.1)	33.6 (9.8)	45.0 (27.3)
DN	17	35.7 (4.2)	20.5 (13.9)	34.7 (4.6)	22.5 (15.9)	41.2 (6.4)	24.5 (20.5)
DT	21	28.8 (5.0)	47.5 (20.7)	32.4 (3.6)	32.2 (17.7)	35.5 (6.5)	41.5 (21.2)
DV	9	26.6 (5.4)	55.9 (17.6)	32.4 (5.4)	36.3 (21.3)	37.2 (7.2)	35.2 (21.2)
ALL	82	30.3 (6.3)		30.3 (6.3)		35.2 (8.2)	

The school mean scores in each scoring method were mostly around the total mean. There was only one school with a mean score with a difference that was more than one standard deviation from the total mean score in each the Analytical marking (DB) and the Comparative Pairs judgements (DL). In the WACE marking all schools were within one standard deviation difference from the total mean score.

The *Rank* columns show the mean for the ranks of the portfolios in each school. The ranking of schools from the three scoring methods were quite different to the scores, partly because of the small sample size in each school. Because the number of students in each school was small, a small discrepancy between scores could create a larger discrepancy in overall rankings. *Figure 4.5* further illustrates the variation of the overall score means obtained from the three scoring processes: Analytical (average of the two assessors for each student), Comparative Pairs, and WACE practical scoring processes.

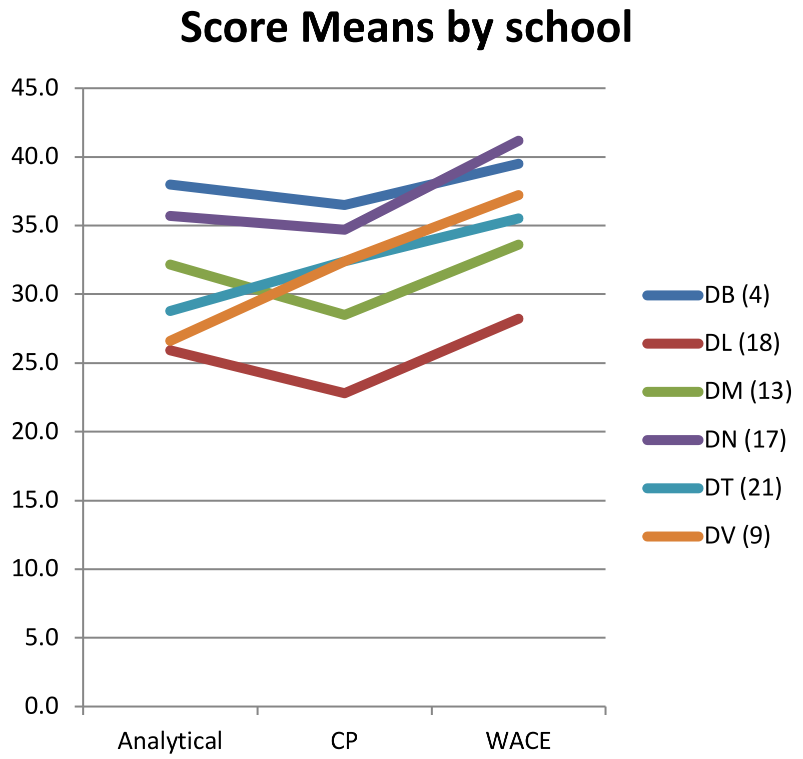


Figure 4.5 Score means by school from each scoring method.

When compared to the other scoring methods, the score means from the Comparative Pairs judgements method were lower for all schools except DT and DV. The score means from the three methods were relatively similar only for school DB. The score means from the Analytical marking and the WACE practical examination marking were quite similar for schools DB, DL, DM and DN but were significantly different in schools DT and DV with differences that were more than one standard deviation. The score means for School DL were consistently the lowest.

The means of the students' overall ranks in each school are depicted in *Figure 4.6*. Apart from school DL, the ranks for the other five schools varied considerably. There was no school with similar ranks between the Analytical marking and either of the two other scoring methods but there was a slight similarity between the Comparative Pairs judgements method and the WACE practical marking ranks especially in schools DN, DV, and DM.

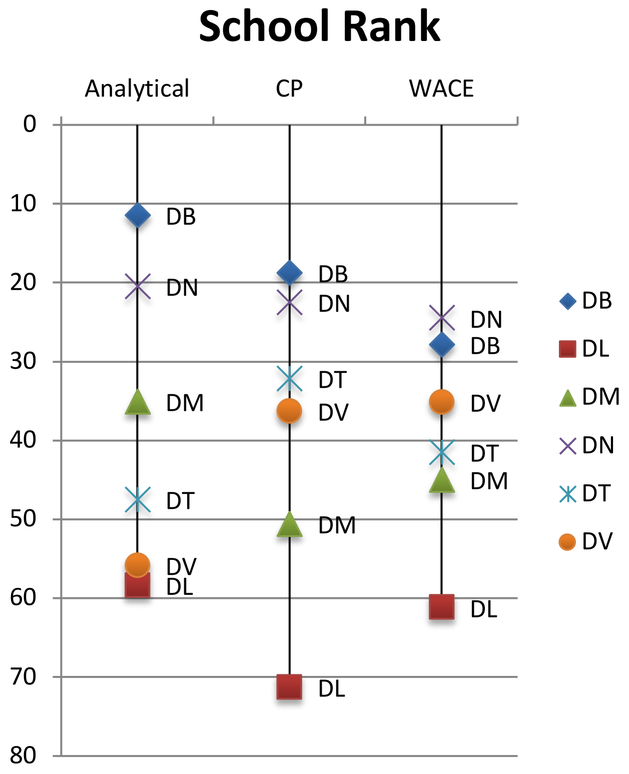


Figure 4.6 Rank means by school from each scoring method.

This presentation of data serves as a preliminary analysis that was aimed to observe possible patterns that might emerge when the schools were compared. In general there was an indication that there could be typical academic characteristics of Design students in each school that might influence student achievement. More detailed analyses are discussed in the next sections.

Comparison between scores from three sources

A correlation analysis was conducted between the scores and rankings that resulted from the three scoring processes. This analysis was done to examine the similarity of the scoring results as part of the validity analysis. The correlation analysis result is shown in Table 4.12 and Table 4.13. In general, there were only low to moderate correlations between the scores from the three methods of scoring. The correlations between scores from the Comparative Pairs judging and the other two methods were moderate and significant, with correlation coefficients of 0.63 and0 0.67 (p<0.01) with Analytical marking and WACE marking consecutively.

Table 4.12
Correlations Coefficients Between Scores from the Three Methods of Scoring

(N=82)	Assessor1	Assessor2	Average	CP	WACE Practical
Assessor1	1	0.53**	0.89**	0.61**	0.55**
Assessor2		1	0.86**	0.48**	0.36**
Average			1	0.63**	0.52**
CP				1	0.67**
WACE Practical					1

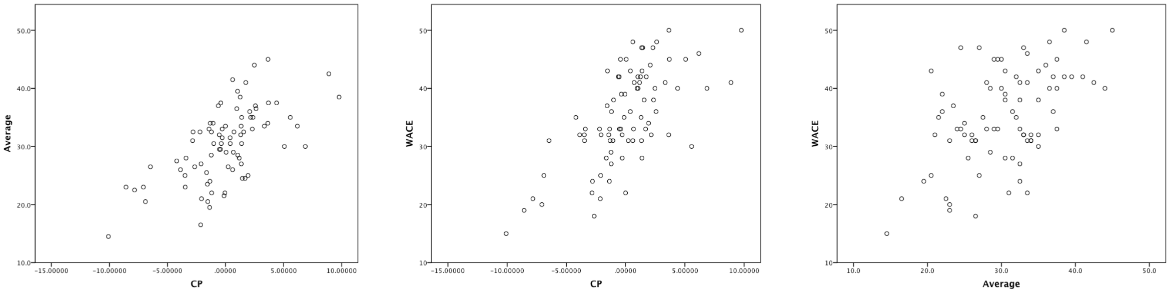
**. Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the .05 level (2-tailed).

Table 4.13
Correlations Coefficients Between Ranks from the Three Methods of Scoring

Rank of	Assessor1	Assessor2	Average	CP	WACE Practical
Assessor1	1	0.54**	0.91**	0.61**	0.51**
Assessor2		1	0.82*	0.48**	0.29**
Average			1	0.63**	0.45**
CP				1	0.63**
WACE Practical					1

**. Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the .05 level (2-tailed).

Even though there were moderate correlations between the scores from Comparative Pairs judgements and the other two scoring methods, there was only a low-to-moderate correlation between the Analytical and WACE scores (r=0.52, p<0.01). The correlations between rankings were relatively similar to the correlations between sets of scores. *Figure 4.7* shows the scatter plot between results from the Comparative Pairs judgements, Analytical marking and WACE marking.



Note: Average = The mean scores from the two Analytical marking assessors

Figure 4.7 Scatter plots between scoring results from the three methods.

The correlation coefficients between the individual assessors' given scores and the WACE scores were 0.55 and 0.36 (p<0.01), consecutively for Assessor 1 and Assessor 2, suggesting that Assessor 1's scores was relatively more in agreement with the WACE results. These correlations were illustrated in the scatter plot graphs in *Figure 4.8* below.

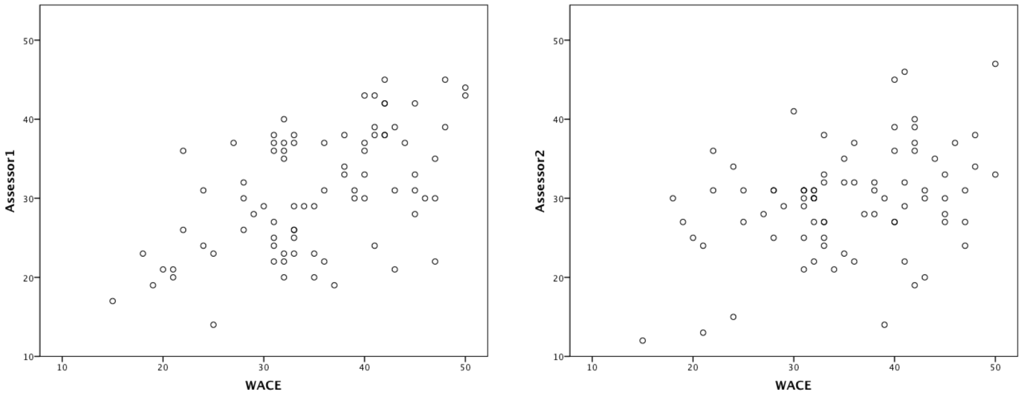


Figure 4.8 Scatter plots between each assessor and WACE.

Similar to that, scores from Assessor 1 were slightly more correlated to the results obtained from the Comparative Pairs judgements than Assessor 2, with correlation coefficients of 0.61 and 0.48 (p<0.01) consecutively. This is illustrated in *Figure 4.9*

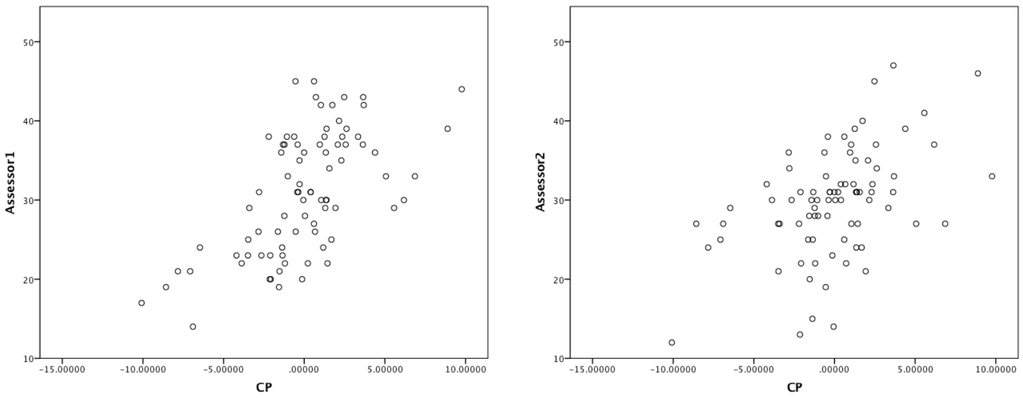


Figure 4.9 Scatter plots between each assessor and CP.

The differences between scores, including between scores obtained from the two assessors, are discussed in the Discrepancy Analysis section.

Discrepancy analysis

Statistics analysis on the results of Comparative Pairs judgements process this far has shown a high reliability in the scores, a moderate correlation with the other scoring methods, and a good fit to a Rasch model. However, these analyses identified that there were several outlier portolios. These were the portfolios that were scored quite differently to the rest when comparing scores from the different sources. Two differences are discussed in this section; the first is based on the difference of the ranks obtained from the Comparative Pairs judgements method and the Analytical marking, the second is based on the misfit analysis obtained from the ACJ system on the Comparative Pairs judgements method. Ranking and scoring data from each assessor in the Analytical marking were presented in addition to the combined Analytical rank and score to better illustrate the similarities and differences in the portfolios that were different to the others. In both analyses the same method is used; patterns that might emerge from the rankings and scorings were discussed, followed by a discussion on assessors' comments from the ACJ system.

Analysis on the discrepancies in the results from the scoring methods repeatedly suggested a conflict between *process* and *product*. In this discussion, *process* describes the complete design process that include almost all aspects assessed in the criteria such as design elements and principles, analysis, innovation, problem solving, skills and knowledge. *Product* refers to the visual aspect of the finished design product, as well as the aesthetic quality of the portfolio. This terminology was decided based on the assessor comments on the ACJ system which suggested that even though unlike the analytical rubric, the holistic criterion only vaguely refers to Criterion 6 of the analytical rubric on Communication and Visual Literacies, the visual aspect on both the portfolios and design products was often influential, especially when the two portfolios being judged were of similar quality. Examples for such comments were “Very close. Production was better in A. B has some good images in final poster” and “Close judgement. A [portfolio A] better resolved”.

Differences between rankings from Comparative Pairs judgements and Analytical marking

Discrepancy analysis between results of the Comparative Pairs judgements and the Analytical marking was conducted based on the ranks obtained from the two scoring methods. Scores obtained from these methods were in different measurement scales. While the Analytical marking resulted in percentage of raw scores, the scores resulting from the Comparative Pairs judgements were in logits. Consequently, the difference between the two scores given to every student was not meaningful nor comparable therefore the ranks obtained from the scores were used instead. Besides, as was discussed in chapter 2, in scoring process there are usually variations in the way assessors distribute the scores. For example, the score of 70% given by assessor A might not represent the same quality as 70% assigned by assessor B, even if they used the same criteria. This is more pronounced in subjective tasks. This section looks more closely into those results in order to establish the cause of the discrepancy between the ranks obtained from the two scoring methods.

Figure 4.10 depicts the distribution of the absolute differences between the ranks generated by the scores obtained from all three scoring processes, while Table 4.14 shows the descriptive statistics of the absolute differences. The absolute differences between the ranks from the Analytical marking and the Comparative Pairs judgements were quite widely spread with differences ranging from 0.0 to 45.5 for 82 students, with a mean of

16.0 and a standard deviation of 12.7. However, when compared to the absolute differences from other pairs of scoring processes such as between WACE and Comparative Pairs or between WACE and Analytical marking processes, this range was the least. The absolute differences of ranking between the WACE marking and the Comparative Pairs judgements were ranging between 0.0 and 60.0, while between the Analytical marking and the WACE marking it was ranging between 0.0 and 63.5. The absolute differences between ranks obtained from the scores given by the assessors in the Analytical marking process were the widest with a range of 0.0 and 76.5, even though these assessors scored the portfolios in the same process.

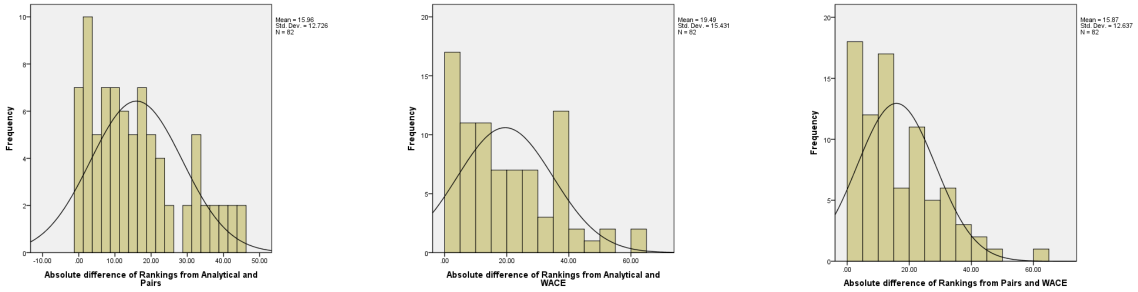


Figure 4.10 Distribution of differences between the rank generated by scores from WACE marking, Analytical marking and Comparative Pairs judgements.

Table 4.14

Descriptive Statistics for Absolute Differences in Ranking Generated by Scores from WACE Marking, Analytical Marking and Comparative Pairs Judgements

Absolute difference Between Ranks	N	Minimum	Maximum	Mean	SD
Analytical - Pairs	82	0.0	45.5	16.0	12.7
Assessor1 - Assessor2	82	0.5	76.5	17.2	15.0
Analytical - WACE	82	0.0	63.5	19.5	15.4
Pairs - WACE	82	0.0	60.0	15.9	12.6

A correlation analysis of the absolute difference in rankings from the three scoring processes, including between assessors in the Analytical marking process, was done to further examine these considerable differences. The correlations are presented in Table 4.15 below. This analysis indicated that there was not much similarity in the difference between rankings from the scoring processes. There was no significant correlation between the difference between the two Analytical marking assessors; and any of the other differences. There were only low but significant correlations between the Analytical-CP differences and the Pairs-WACE differences, and the latter with Analytical-WACE differences, with correlation coefficients of 0.36 (p<0.01) for both. This suggests that there was a weak possibility the different scoring process and type of criteria could be one of the factors that created difference in the scoring results.

Table 4.15

Correlations Between Absolute Differences in Ranking Generated by Scores from WACE Marking, Analytical Marking and Comparative Pairs Judgements

	Analytical - Pairs	Assessor1-Assessor2	Analytical - WACE	Pairs - WACE
Analytical - Pairs	1	0.11	0.36**	0.03
Assessor1-Assessor2		1	0.02	0.10
Analytical - WACE			1	0.36**
Pairs - WACE				1

**. Correlation is significant at the .01 level (2-tailed).

The lack of strong correlations between the absolute differences of ranks obtained from the three scoring processes further implied that there was no specific consistent procedural reason for the large differences. It indicated that the absolute differences were not caused by differences between scoring methods, which were the difference in criteria, scoring media (i.e., original paper portfolio or digital portfolio), and calculations to obtain the final scores. Consequently, it indicated that the differences were most likely to be caused by factors such as the portfolio quality (e.g., the length, the focus, the quality of the components), the quality of the scoring criteria (e.g., the range of scores, semantics) or the assessors' preference. It should also be noted that statistically the small sample size could also cause the differences between the distance between scores and the distance between ranks, amplifying the distance between scores during the ranking process. This effect is illustrated in a later paragraph.

Because there was no indication that the differences in the rankings were caused by procedural factors in the three scoring processes, the next step was to examine other factors that could cause the difference in the rankings. Portfolios with more than 2 standard deviations difference from the mean of the absolute difference between ranks obtained from the two scoring methods were analysed to investigate the possible main reasons for the difference such as the quality of the portfolios, the assessor's personal preference, or technical problems. In the Design course there were four out of 82 portfolios (4.9%) with such large difference. Three from those four portfolios were from school DT. However, this might not be conclusive as school DT had the largest number of portfolios compared to any other schools. Table 4.16 shows the ranks and scores for the four portfolios. For these four portfolios, the ranks obtained from all three scoring methods were very different and as the previous correlation analysis indicated, aside from some similarities in raw scores there was no obvious pattern emerging from the differences.

Table 4.16
Portfolios with More than 2 SD Difference in Ranking in Design Course

ID	Rank					Score (%)				
	Analytical			CP	WACE	Analytical			CP (rescaled)	WACE
	Ave	A1	A2			Ave	A1	A2		
DT920	66.5	70.5	59	23	6	49	44	54	66	94
DT921	66.5	59.5	70	21	47	49	50	48	66	66
DT922	64.5	49.5	79.5	19	43	50	58	42	68	68
DV909	45.5	35	55	3	28	60	66	54	86	80

With regards to raw scores, there was agreement for portfolios DT921 and DT922, between the Comparative Pairs judgements and the WACE marking that produced the same scores for both portfolios, which were 66% and 68% consecutively. Scores from the Analytical marking were significantly lower, which were 49% and 50% for DT921 and DT922 respectively. For portfolio DT921, each assessor gave relatively similar scores of 50% and 48% while for portfolio DT922, the assessors gave it scores of 58% and 42% respectively.

When the score distribution in each scoring method was considered and the scores were converted into ranks, these slight differences between scores from the different methods of scoring, and different assessors in the Analytical marking, were magnified. For example, for DT921 and DT922 whose scores obtained from the Comparative Pairs judgement were exactly the same as the scores obtained from the WACE practical marking, the ranks were very different. The Comparative Pairs judgements placed DT921 on the 21st and DT922 on the 19th while the WACE practical marking placed them on the 47th and 43rd consecutively.

While the scores suggested that the works were scored similarly, the ranks were in different quartiles.

Portfolios DT920, DT921 and DT922 were from the same school, DT. The ranking of these portfolios showed that in Comparative Pairs judgements they were considered to be of similar qualities as they were ranked as the 23rd, 21st, and 19th consecutively. These ranks were quite close together and were all within the first quartile of the ranking. In Analytical marking, all three were ranked as the 66th, 66th, and 64th consecutively, which were also close together, but within the last quartile. Within each of these scoring methods, these portfolios were considered of similar quality but between the two methods they were judged to be very different. Even though the correlation analysis on the differences in the rankings from the three methods did not indicate procedural differences, this pattern suggested that there was still a possibility that for a few portfolios the procedural difference between the two scoring methods caused a difference in the resulting ranks. The small number of such portfolios could be the reason the correlation analysis did not suggest significant association between differences in rankings from the three scoring methods. Therefore, in the following discussion, the type of criteria, which was the procedural difference between the two methods that could affect the difference in the rankings, is still considered. Beside the type of criteria, the difference between the ranks obtained from the Comparative Pairs judgements and the Analytical marking for portfolios DT920, 921, and 923 could also be caused by the differences between the assessors' personal preference, components of the portfolios, or a combination of these factors.

In terms of the type of criteria used, in the Comparative Pairs judgements it was a holistic criterion while in the Analytical marking, as well as the WACE marking, it was an analytical marking rubric. If the type of criteria was the cause of the ranking difference in these three portfolios, then the rankings from the scoring processes in which the analytical marking rubric was used should be relatively similar. Rankings from each Analytical marking assessors and WACE all varied for these portfolios. The WACE marking result placed DT920 as the 6th, which was very different to the other two methods, and DT921 and DT922 as the 47th and 43rd consecutively, which was relatively more similar to the

Analytical marking. *Figure 4.12* and *Figure 4.11* below illustrate the rankings and scores for these four portfolios obtained from the three scoring methods.

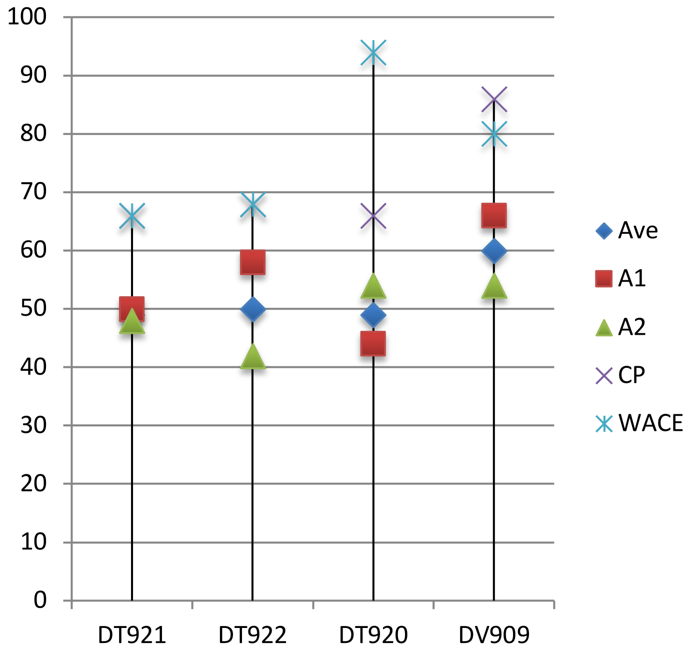


Figure 4.11 Scores for portfolios DT920, DT921, DT922, and DV909.

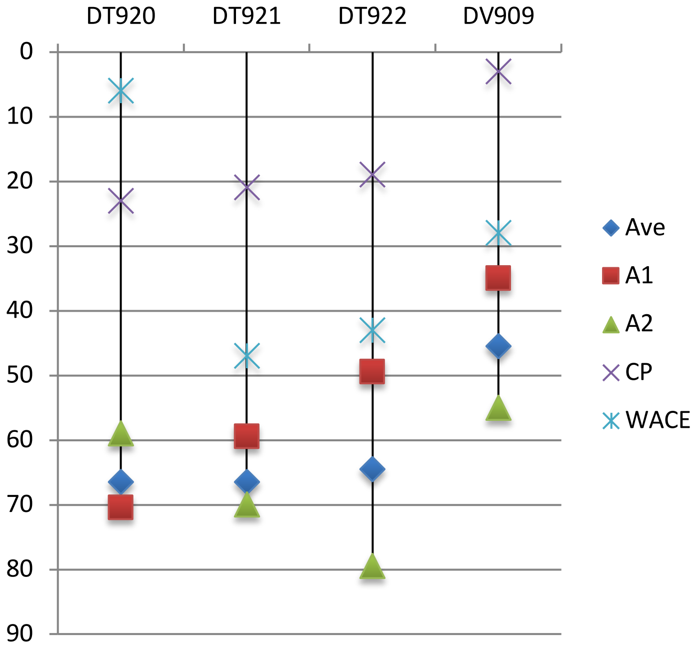


Figure 4.12 Rankings for portfolios DT920, DT921, DT922, and DV909.

One of the main differences between the WACE marking and the Analytical marking processes was the form of the portfolios used in the marking processes. In the WACE marking process, the assessors marked the original paper portfolio while in the Analytical marking process, the assessors marked the digital version. Another difference was in the way the final score was obtained in each process, especially when the assessors assigned very different scores. In Comparative Pairs judgements this was not a factor because of the way the ACJ system created the pairings and calculated the parameters, the judgements process in which the assessors did not assign a score, and the multiple number of assessors which cancelled out assessors' bias. In the Analytical marking process regardless whether the difference between scores assigned by the two assessors was large or not, the final score was the mean of the two scores. In the WACE marking process, when there was a large difference between scores given by the two assessors, the scores were discussed to decide on a score. This could mean the WACE final score of such portfolio was the score from one assessor as opposed to the mean of the two scores. As a result, bias could be one of the factors that caused the differences between the WACE ranking with the other two rankings.

There were several similarities among rankings and scores from the three methods but there was no apparent pattern that could be used to explain the large difference of rankings from the Comparative Pairs judgements and the Analytical marking methods. A possible source of such difference was assessors' subjectivity and the way the features of the portfolios might attract or repulse particular assessors. Therefore a closer look into possible assessors' preference and portfolio characteristics through assessors' comments for the portfolios were the next step.

Not much information could be found from a more detailed look into each of the Analytical marking assessors' result for DT920, while for DT921 and DT922 there was a vague pattern of Assessor 1 giving higher scores, especially in Criterion 3 onwards. This might not indicate much, considering those criteria had a wider range than the first two criteria, which means that they had a wider spread, with score ranges from zero to ten for C3 to C5 and zero to eight for C6. The first two criteria had a more narrow spread with a score range from zero to six each; hence it was more probable that the two assessors gave the same score. Another explanation could also be a disparity in the two assessors' preferences. From the main project it was found that Assessor 1 had a tendency to focus more on the process while Assessor 2 was more particular about the quality of the product (Newhouse et al., 2012).

The Comparative Pairs assessors considered these three portfolios to be quite strong products with comments such as *good production skills shown*, *final product is detailed*, *final design works meet industrial code and standards*. These portfolios were not considered to show a strong design process as shown by comments such as *design process has some gaps*, *Process is more complete and final is stronger in A* (the work compared to this work), *design process is lacking in development of ideas*. A contrast between the quality of the design process and the product could be one of the reasons of the difference between results from the Comparative Pairs judgement and the Analytical marking.

DV909 was not very different to the other three portfolios. Ranks from the three scoring methods were all different with Comparative Pairs assessors ranking the portfolio as the 3rd, Analytical marking assessors ranking it as the 45th, and the WACE markers as the 28th. The difference between the two Analytical marking assessors was also the same, with Assessor 1 giving higher scores for Criterion 3 onwards. Two pages from the DV909 digital portfolio were blank, however, no assessor commented on this and almost all Comparative Pairs assessors' comments were positive and indicating strong design process and product. In scoring processes that used an analytical marking rubric, there was a possibility that when there were missing components such as in this portfolio, regardless of whether it was a scanning mistake during digitisation process or a mistake from the student's side, assessors deducted scores from one or more criteria. Since in the Comparative Pairs judgements the assessors did not assign a score, there was no deduction of scores either. In the case of portfolio DV909 it was possible that the rank from the Comparative Pairs judgements was the highest while there were variations in the ranks from the other processes because of this. As long as the portfolio was stronger than the comparison portfolio based on the holistic criterion, it would still be judged as the winner. In Comparative Pairs judgements, minor mistakes and weaknesses could be overlooked as long as the overall criterion was met. Therefore, the Comparative Pairs judgements with a holistic criterion might not be suitable for assessment tasks that consist of numerous detailed skill components.

Overall, an examination on the four portfolios with a large difference between rankings obtained from the Comparative Pairs judgements and the Analytical marking did not reveal conclusive causes to the difference. The investigation did not suggest that the difference was caused by procedural differences between methods of scoring, therefore it could be caused by subjective factors such as assessors' preference and portfolio characteristics. It is possible that there was an exception on portfolios with strong qualities, either it was very good or very bad, that differences in scoring processes and criteria could cause a large difference in rankings because in Comparative Pairs judgements small mistakes might be overlooked.

Comparative Pairs misfits

The ACJ system provided data on misfits for both the assessors and portfolios. Based on the weighted mean square values (wms) on the portfolios, there were six portfolios (7.32%) with a difference above two standard deviations to the average value. These misfits were portfolios that were judged differently by different assessors, and Pollitt (2012a, 2012b) suggested that because the assessors could not agree on the ranking of such portfolios, the portfolios should be examined more closely.

This difference indicated that there could be a disparity between the quality of the portfolios and the criterion used to judge them. The disparity could be from the assessors' side or the students' side, or both. Several factors could be the reason of this gap, for example assessors' personal preference, a lack of the students' understanding on WACE criteria that were used to develop the holistic criterion, the students' inability to communicate their design, or missing or unaddressed rubric components. Table 4.17 displays the six portfolio results from the different scoring processes.

Table 4.17

Ranks for Portfolios with Weighted Mean Square (wms) More than 2 SD Difference in Design Course

ID	Rank					Score				
	Analytical			CP	WACE	Analytical			CP	WACE
	A1	A2	Ave			A1	A2	Ave		
DB906	11	28	18.5	24	15	39	31	35	33	43
DL903	59	75	71	73	61	25	21	23	24	31
DM910	20	51	33	56	70	37	28	32.5	28	27
DN902	48	4	18.5	5	65	29	41	35	41	30
DN926	7	20	10	8	10.5	42	33	37.5	37	45
DV907	34	55	45.5	6	10.5	33	27	30	40	45

Unlike in the analysis of differences between rankings from the Comparative Pairs judgements and the Analytical marking, the six portfolios with wms more than two standard deviations difference did not show any particular similarities in their scores and rankings. It should be noted that the parameter calculation for the Comparative Pairs judgements was not a simple linear function but instead iterations of numerous probability functions dependent on which portfolios they were compared with. Therefore, even though DB906 and DL903 had the same number of wins and losses, it was possible that they could have different ranks, which in this case DB906 was ranked at the 24th and DL903 was at the 73rd. *Figure 4.13* and *Figure 4.14* illustrate the ranks and scores for the six portfolios.

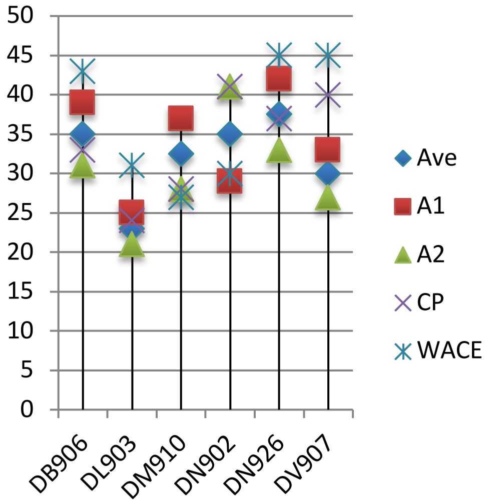


Figure 4.13 Score for portfolios with wms more than 2 SD difference in Design course.

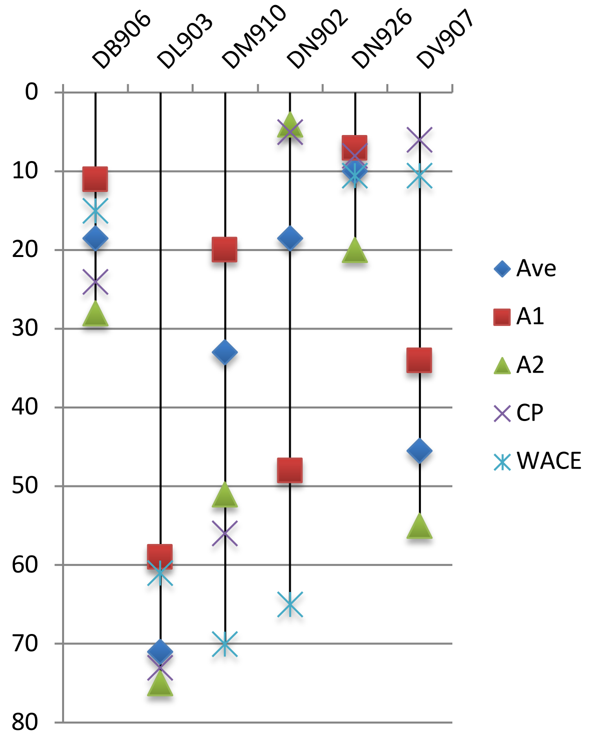


Figure 4.14 Ranks for portfolios with wms more than 2 SD difference in Design course.

Because there were similarities among portfolios DB906, DL903 and DN926 in terms of the distribution of rankings from the three methods, these portfolios were discussed together. These three portfolios were ranked quite closely together in the three scoring methods. Portfolios DM910, DN902 and DV907, conversely, were distributed across methods and thus also discussed together.

Portfolios DB906, DL903 and DN926

These portfolios were ranked quite closely together in the three scoring methods. The ranks from Assessor 1 in the Analytical marking were consistently the highest and the ranks from Assessor 2 were consistently the lowest for these three portfolios.

Portfolio DB906

For portfolio DB906, the ranks and scores obtained from the three methods were not too different, which showed agreement across assessors and methods, with ranks ranging from the 11th from Assessor 1 and the 28th from Assessor 2, both in Analytical marking with the Comparative Pairs judgements process placed it on the 24th. The wms of this portfolio, however, was quite different to the mean. While the large difference in the wms value indicated that one or more assessors judged this portfolio differently to the rest, rankings from the other methods showed that the portfolio was ranked quite similarly. This, in turn, could indicate that the holistic criterion resulted in inconsistency when it was used to judge this portfolio.

Assessors' comments on the ACJ system varied with comments that indicated average quality such as *production is average, idea is not original, process has some gaps – only one idea explored to resolution (repetitive)* and *final images needed modifications to create a more realistic edge* as well as comments that indicated superior quality such as *Well analysed considering most relevant information* and *Analyses development and final works show better understanding* (than the other portfolio) *and application of elements and principles of design*. In general the comments suggested that the components of the portfolio that showed design process were of average to high quality but the components that showed the design product was only of average quality. These varied comments were parallel with the large wms deviation of this portfolio. They also further supported the earlier proposition that several assessors were more concerned about evidence of design process while others were more attracted to the quality of the design product. These comments also suggested that there were assessors who judged the portfolio holistically and considered the components on both the design process and product.

The large wms deviation and the agreement between assessors in Analytical marking as well as among methods were contradictory. This contrast suggested that there were components of this portfolio that did not quite match the holistic criterion or that there was disagreement among Comparative Pairs assessors on how to judge this portfolio based on the holistic criterion.

Portfolio DL903

Similar to DB906, there was agreement across methods for DL903, however DL903 was ranked quite low in the rankings with the highest of the 59th as ranked by Assessor 1, the lowest of the 75th by Assessor 2 in the Analytical marking, and the Comparative Pairs judgements ranking of the 73rd. As expected, assessors' comments for this portfolio were also varied with comments such as *judgement was a little difficult when they are equally poor but across two different disciplines*, *Heavy investigation and analysis output however poor translation of concepts into final flyer. Process steps not extensive. Typography poor*, as well as *Final design shows great design composition* and *Shows good photographic skills and techniques*. These comments highlighted the different components of the portfolio that were considered more important than the rest by several assessors. While in scoring processes that use an analytical scoring rubric the distribution of scores across different assessed skills was set and therefore there was a degree of uniformity across assessors and across skills, in Comparative Pairs judgements assessors' personal preference could affect their judgements and create a difference in judgements compared to other assessors. Potentially this is more apparent in tasks that assess more skills.

Portfolio DN926

DN926 was also similar to DB906 and DL903 in terms of rankings. The ranks from Assessor 1 was also the highest for this portfolio, placing it on the seventh, and from Assessor 2 the lowest on the 20th. The Comparative Pairs judgements placed this portfolio on the eighth. Unlike the other two portfolios, however, there was no comment that could explain the reason the wms value was too different to the mean. All comments for this portfolio were positive and when the assessors decided that the portfolio that was compared to this portfolio was the winner it was mostly because the other portfolio was of a better quality. Comments from the assessors included *Clever idea simple execution and effective. Design development to the point and original*, *Excellent process. Good exploration of ideas. Great analysis of own work and others. Strong links to TA*, *Final design includes audience's input and therefore more effective*. *Good use of elements and principles of design*, while comments on the comparison were such as *A* (the other portfolio) *is slightly deeper in context and thinking* and *A* (DN926) *shows a more consistent design process and flow of ideas. B* (the other portfolio) *is stronger than others in this group* that imply that the judgements were on portfolios with similar qualities.

Portfolios DM910, DN926 and DV907

These portfolios were ranked very differently in the three scoring methods. This variation in rankings across methods corresponds to the large wms deviation in these portfolios. There was also no apparent pattern among the ranks, unlike in the previous three portfolios.

Portfolio DM910

Portfolio DM910 had a range of ranks from the 20th from Assessor 1 scores to the 70th from the WACE scores. The Comparative Pairs judgements method placed this portfolio on the 56th which was relatively close to the rank from Assessor 2. Contradictory comments for this portfolio suggested a conflict between emphasising judgements on the quality of design process or the product. There were comments on good design process such as *Good experimentation. Good analysis. Quality production skills demonstrated*, as well as comments on mixed portfolio quality such as *Design works are imaginative. Needs more consideration on ergonomics and practicality* and *Excellent concepts of the two final ideas. Logo ideas could show more design development*, and final presentation comments such as *Low level sketching and perspective/dimensional aspects" and "Lack of highly polished final presentation brief addressed*. This variation of comments was parallel to the large wms difference on this portfolio.

Portfolio DN902

The ranks for portfolio DN902 ranged widely between the 4th (Assessor 2) and the 65th (WACE). The ranking from the Comparative Pairs judgements placed this portfolio on the 5th, very similar to the ranking from Assessor 2 from the Analytical marking. Unlike the other five portfolios with large wms difference, in the Analytical marking this portfolio had a much higher Assessor 1 rank than Assessor 2 rank, highlighting the possibility that for some portfolios the two Analytical assessors' judgements could be very different even though they were using an analytical marking rubric.

Assessors comments from the ACJ system did not provide information that could explain the large wms difference for this portfolio. Comments on judgements suggested that when this portfolio lost it was because the other portfolios were slightly better, for example *Very close, a hard one to judge. B* (the other portfolio) *had a more detailed response to the brief and a more mature finish*. Comments on this portfolio were mostly positive and quite detailed, for example *Design works show better understanding of element and principles of design. Works also involve good user research* and *Strong design concept. Excellent typographic arrangement. Design shows great layout skills. Communicates message across well*. The only problem with this portfolio mentioned in the comments was the number of class notes included, as one assessor commented *Good example of design process. Strong justification of design intent and decisions made final production is high. Too many in class notes filling pages – is not evidence of understanding*.

The large wms difference in the Comparative Pairs judgements result and the wide range of ranks for this portfolio from all three methods indicated that the assessors' judgements widely varied regardless of the processes, scoring media and criteria used. Comments from the ACJ system did not provide information on the quality or components of this portfolio that could cause the varied judgements; the comments only indicated that the quality of this portfolio was high, which was contrary to the WACE ranking.

Portfolio DV907

The rankings for portfolio DV907 from the three scoring processes were varied with the Comparative Pairs judgements placed the portfolio the highest on the sixth and Assessor 2 from the Analytical marking placed it at the lowest on the 55th. The WACE rank was quite close to the Comparative Pairs rank at the 10.5th. Assessors' comments from the ACJ system for this portfolio were all positive with no weakness mentioned. There was no information from the comments that could refer to the large wms difference in this portfolio but the ranks from the three methods were widely spread. These comments were relatively detailed and suggested that the assessors were familiar with the analytical marking rubric, for example *Good exploration of alternative ideas. Good justification and analysis of own decisions made. Production is good. Design process well documented* and

Employs a range of skills. Strong final concept. Applies simple graphic elements that communicate to the audience.

For portfolios DN902 and DV907 there were no comments that could provide information on possible reasons for the large wms difference but the variety in ranks from the three methods was parallel to the wms difference.

Assessor Interview

After both scoring processes were concluded, the assessors were asked to give their opinions on the scoring processes, the online tools used, and the quality of the work submitted by the students. Eight of the 10 assessors sent back their responses through email. These assessor demographic data are as shown in Table 4.18. The interview consisted of five demographic questions and twelve questions pertaining to the assessors' experience in the scoring processes. These questions are presented in Appendix E.

Table 4.18
Design Assessor Demographic Data

Assessor	Age Group	Teaching experience (Years)	Teaching Design (Years)	Teach Stage 3 Design in 2011	WACE marker
A	30-40	7	7	No	Yes
B	>40	25	25	No	No
C	30-40	7	4	No	Yes
D	30-40	6	6	No	No
E	30-40	13	13	No	Yes
F	>40	17	0	No	Yes
G	>40	20	8	No	No

All eight assessor-respondents had more than five years of teaching experience, and all but one assessor taught courses related to the Design course, with only one of them having never taught the Stage 3 Design course. In 2011, none of the eight assessors taught Stage 3 Design, but half of them were involved in the 2011 WACE marking.

The rest of the questions in the interview were designed to gather the assessors' opinion on the quality of the student work, the scoring processes and the suggestions they had regarding the whole assessing experience. In this study, the assessor interview was used to provide information on the assessors' experience that might reduce the validity of the scoring result. Therefore, only responses that pertained to issues surrounding reliability and validity are discussed in the next section.

Reliability of Scores

As was discussed in Chapter 3, the ACJ system was designed so that it systematically created judgement rounds that would gradually become finer and finer in pairing student work. As this process was being done, the reliability of the judgement also became higher, mostly because of the combination of the gradual increase in the number of judgements that consequently increased the cancelling out of the differences between judges and the gradual improvement in the fine-tuning of the pairings. Once the reliability coefficient reached the intended value, which was also when more judgements did not increase the reliability much, the judgement process was concluded. In Design, this happened when the reliability coefficient reached 0.941. This high reliability level reflected both the inter-rater reliability and internal reliability, as calculated by the ACJ system (Kimbell, 2008).

Because of this characteristic, the Comparative Pairs judgements method was likely to reach a high reliability coefficient, unless the misfits were too extreme. The reliability coefficients of the scoring methods were as shown in Table 4.19. There was no reliability analysis available for the WACE result. The WACE scores were obtained from double-blind marking and reconciliation between markers. The Comparative Pairs judgement reliability coefficient was obtained from the analysis generated by the ACJ system using a statistic analysis similar to Cronbach's alpha coefficient.

Table 4.19

Reliability Coefficient of the Analytical Marking Results

Method of marking		Internal reliability
Analytical marking:	Assessor 1	0.953
	Assessor 2	0.950
	Average	0.962
Comparative Pairs marking		0.941
WACE Examination		n/a

The high internal reliability specified by the Cronbach's Alpha coefficients obtained from the SPSS software for the Analytical marking represented the internal reliability of the criteria. These reliability coefficients indicated that there was an overall agreement among the criteria in the rubric. The inter-rater reliability was represented in the correlation between assessors. However, even though the internal reliability for the Analytical marking was high, the correlation between scores from the Analytical marking for the two assessors was only moderate, with a correlation coefficient of 0.53 and 0.54 ($p<0.01$) respectively for score and rank, as was shown in Table 4.12. These coefficients indicated that there were only moderate correlations between assessors in the Analytical marking.

Parallel to this, as was discussed earlier in this chapter, 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 main project it was found that for some criteria, the lowest and top most ends of the score range were not used by the assessors, especially by Assessor 2. In contrast, the reliability coefficient of the Comparative Pairs judgement, which represented both the internal reliability, or internal consistency in judgement, and the inter-rater reliability, was high.

Validity of Assessment

Three points of reference are used to discuss the validity of the Comparative Pairs judgements. The first is from the reliability of the result of judgement, then from the way the result was compared from results from the other scoring methods, and lastly, from the issues that might threaten the validity of the result as were disclosed by the assessors in the interview.

Reliability of scores supports validity

The result from the Comparative Pairs judgements had a high reliability coefficient, therefore the threat from the lack of both the internal and the inter-rater consistency can be regarded as low. As Pollitt (cited in Kimbell et al., 2009, p. 79) posited, in the Comparative Pairs judgements method, variation in both the absolute standard and the weightings did not influence the validity of Comparative Pairs judgement result. One factor that might have an effect on the validity of this result was the variation in portfolio qualities that were judged to be the winner or the loser mostly based on the judges' personal preference. Considering that the internal reliability in the Comparative Pairs judgement result was 0.941, which was high, there was a high confidence in the consistency of the judgements. The assessor misfit statistics, as shown in Table 4.7 also did not indicate extreme inconsistency in judgements.

In contrast, for the Analytical marking, the internal reliability was high but the inter-rater consistency was only moderate. This indicated that even though the criteria measured the same set of skills, there was inconsistency in how the assessors used the rubric, which in turn lowered the reliability of the scores, and consequently reduced the validity of the scoring method.

In both scoring processes, only experienced assessors were selected. This was aimed to avoid differences among assessors that were caused by lack of experience. Technical help was also provided in both processes to avoid disturbance by technical problems such as difficulties in accessing the interface. For the Comparative Pairs judgement the holistic criterion was discussed together by most assessors based on the WACE examination criteria. This was aimed to avoid differences in understanding the holistic criterion. These efforts were taken as a precaution to limit the factors that could potentially compromise the validity of the result.

Comparison with results from other scoring methods

Comparability with results from other scoring methods is a measure of validity. Comparison with results from other methods indicates the generalisability of the result from the Comparative Pairs judgements method. In this study, the result from the Comparative Pairs judgements method was compared to the results from the WACE marking and the Analytical marking, as well as to how the result from the Analytical marking correlated to the result from the WACE marking.

The WACE marking result was obtained from a rigorous analytical marking process conducted by the Curriculum Council. The process included double-blind marking by two assessors and reconciliation of scores when there was a significant difference in the two scores. If we assume that the official WACE scores for the original portfolios are a valid measure of what is intended by the assessment then if other sources of scores correlate strongly with the WACE scores this indicates that they are likely to be measuring the same thing. As a comparison, the recommended agreement between assessors is 70% or above (Brown, Glasswell, & Harland, 2004; Jonsson & Svingby, 2007; Stemler, 2004).

As discussed in an earlier section, the correlations between scoring methods in Table 4.12 indicated that results from the Comparative Pairs marking was significantly and moderately correlated with results from both the Analytical and the WACE practical markings with correlation coefficients of 0.63 and 0.67 consecutively for the scores and 0.63 for both rankings, which were relatively close to the suggested agreement level of 70%.

As a comparison, the correlation between assessors in the Analytical marking was relatively lower, with a correlation coefficient of 0.53 ($p<0.01$). The correlations between individual Analytical marking assessors and the WACE scores were 0.55 for Assessor 1 and 0.36 for Assessor 2 ($p<0.01$). This suggested that the scores from the Comparative Pairs judgements were relatively more similar to the WACE scores than the Analytical marking scores. Considering the WACE and the Analytical markings both used the same method and marking rubric while the Comparative Pairs judgements used a different method, the slight differences in the correlations was considered significant.

Furthermore, even though the internal reliability of the scores from both the Comparative Pairs judgements and the Analytical marking was similarly high, the low correlations between assessors lowered the confidence on the validity of the result of marking. Averaging the results from the two assessors did moderate the results but that still did not quite bring the confidence level in the results to the same level as the results from Comparative Pairs judgements.

Validity issues emerging from the assessor interviews

Regarding the quality of the digital representation of student work in Design, most assessors considered the quality to be adequate, aside from a few portfolios that were not scanned well, such as pages with pencil sketches. Examples of these comments are: *Graphic Portfolios show clear representation of students' work* and: *As the originals are 2D, this translated well into digital format*. An assessor made a distinction between the required qualities for analytical marking and Comparative Pairs judgements processes: *For subtle difference and qualitative assessment the paper gives more information. For ranking the digital folios are as good*, indicating that even though in some portfolios the details lacked clarity, he did not consider this to affect his judgements in the Comparative Pairs process. Several assessors considered navigating the original portfolio to be more convenient than the pdf version with comments such as: *Paper copies seem to be a little easier to flick back and forth and there is definitely a need for large monitors to view the work adequately* and: *Zooming in allows you to see detail but blocks out other aspects e.g., you may be able to read an explanation but can't see the diagram associated with the explanation – this requires the marker to move around the screen using the mouse, rather than being able to 'flick' between the written and visual at a glance*. Even though the assessors reported several problems with the quality of the pdf and the inconvenience of the monitor size, all of them were satisfied with both digital scoring processes and stated: *there was no difficulty in understanding the students' abilities and performance levels* and *it was no different than viewing the printed versions*. From the assessors' report on the quality of the digital portfolio, it was likely that this did not overly compromise the validity of the result from both the Comparative Pairs judgements and the Analytical marking.

All assessors considered both the ACJ system for the Comparative Pairs judgement and the Filemaker Pro database for the Analytical marking to be easy and convenient to use.

One assessor reported that for Analytical marking the original portfolio would be better because the original portfolio gave more information necessary for it, while for the Comparative Pairs judgements the digital portfolio would be as good as the original. The Comparative Pairs judging was reported to be difficult to do when the types of the portfolio compared were different, for example when one was a technical graphic portfolio and the other was a photography portfolio. There was also reported difficulty in judging two portfolios that were equal in quality.

Regarding the scoring processes and the criteria used in those processes, there was a mixed response from the assessors. In general, the Comparative Pairs judgements method was considered to be more accurate, more straightforward, and more objective. On the other hand, the Analytical marking was considered to allow for a more careful and accurate scoring. Assessors who preferred the holistic criterion considered the criterion made the judgements easier and more accurate because the criterion was easy to remember and also because it was easier to compare portfolios than to assign scores based on an analytical marking rubric. Such comments were: *I found the pairs marking less demanding than analytical marking. I didn't need to hold standards in my head*, *Folios are judged multiple times by many different markers which helps avoid the problems of inconsistencies*, and: *Different interpretations of the marking key often led to discrepancies and this was eliminated through this Comparative Pairs marking process*.

Assessors who preferred to use a rubric considered it to be more specific, accurate and accountable with comments such as: *I would prefer analytical marking as this allows me to analyse and judge one design work at a time. This focus is more detailed and accurate – for me*" and

> Coming from an old school approach, I still prefer the analytical. Saying that, the judgements I was able to make were sound and justifiable. I would have preferred a hybrid of the two, with more than 1 criterion for assessment of the digital form. I'm sure students, teachers & parents would also be a bit miffed if they knew that the judgement of 50% of the students work was based on a single sentence.

About the factors that might influence their judgement, the assessors reported several sources. Some of them were involved in the WACE marking, therefore they may have already seen the original portfolio and thought that might affect their judgement. Several assessors mentioned that there were portfolios that were: *heavily reliant on teacher-based notes or very rigid templates it was hard to discern what was really student understanding and what was padding*. There were also: *closed projects that limited strong students to push boundaries and achieve top results*. One teacher considered his previous study in Bachelor and Masters in Design and both teaching and professional experiences as influential factors in his judgement. Another teacher put his own expectation of standard to be affecting his judgement, even though he tried to follow the assessment criteria.

Summary

This chapter presented the analysis of data from the Design course, starting with the constraints related to the assessment task which included time constraints, the types of portfolio materials that made the scanning process more challenging, and technical limitations caused by the size of the digital portfolios. Data from the three scoring methods was presented next.

Scoring data from the Analytical marking showed good agreement among criteria and a moderate correlation between assessors. There were only three portfolios that were marked too differently by the two assessors. Compared to the WACE scores, the Analytical marking scores were lower and the range was narrower, illustrating the underutilisation of the lowest and highest scores. Comparative Pairs judgements data from the ACJ system demonstrated good agreement among assessors, with only 4.6% judgements considered inconsistent and 7% of the portfolios were identified to have different judgements. Comparisons between scores from the three sources showed moderate correlations, with Analytical assessor 2 having only low correlations with either the Comparative Pairs judgements and the WACE marking.

The possibility that ranking difference between the three methods was caused by procedural differences such as the difference in criteria or scoring media could be overlooked because there was no strong correlation between the differences in ranks. Discrepancy analysis of portfolios which was conducted on portfolios that were ranked too differently between the Analytical marking and the Comparative Pairs judgements indicated that the differences more likely to be caused by subjective factors such as assessors' preference and the effect of portfolio quality on judgement. Assessors' notes on ACJ system related to judgements of the portfolios with high misfit statistics also indicated the same subjective factors, with a possibility that Design assessors had a tendency to prefer either portfolios that showed stronger process or product.

The reliability coefficient reported in the ACJ system for the Comparative Pairs scores represented both the internal reliability and the inter-rater reliability of the scores. The internal reliability of the scores from the Analytical marking and the Comparative Pairs judgements was similarly high, however, the inter-rater reliability between assessors as indicated by the correlation coefficients between the two assessors in the Analytical marker was only moderate.

The assessor interview indicated that the quality of the digital representation was sufficient. Even though several assessors preferred to navigate through a paper portfolio, they found that the online scoring processes to be easy to use and the Comparative Pairs judgements were easy to make, except when the paired portfolios were of similar quality or on different course contexts. The analysis of data from the Visual Arts course is presented in Chapter 5.

CHAPTER 5
FINDINGS FROM THE ANALYSIS OF DATA – VISUAL ARTS

This chapter presents the result from an analysis of the data collected for the Visual Arts course assessment in a similar manner to Chapter 4. This will allow comparisons between the two sets of results in the next chapter. Therefore, this chapter starts with a description on the artworks submitted followed by a presentation of the results of an analysis of the data from the Comparative Pairs judgements method and the Analytical marking of the digital representation of the artworks, as well as a comparison with the data from the official WACE analytical marking. Then a discrepancy analysis between the scores from the different sources is discussed. Next an analysis of the assessor interview data is presented, combined with an analysis of the reliability of the scores, and validity of the Comparative Pairs judgements. The purpose of this presentation of results is to provide findings with which to address the research questions in a later chapter.

Student Work

The practical task for the Visual Arts course WACE examination was submitted as a finished artwork accompanied by an artist statement and an installation photograph. Students' completed works were delivered to the Curriculum Council at a designated location, where the Curriculum Council staff received, labelled, catalogued, and arranged them for the WACE examination marking. However, because of the space limitation, many of these artworks were not installed, in the main they were stored in labelled boxes. For the main project in which this study was located, the participating students' works were digitised in the forms of photographs and videos. This digitised version of student work was used in both the Analytical marking and Comparative Pairs judgements processes while for the WACE examination marking process the examiners used the original artwork. In this section, factors that could affect the scoring data from the nature of student work, the digitisation process, and the technical limitations are discussed.

Nature of student work

The practical task for the Visual Arts course WACE examination was submitted as a resolved artwork that could be two dimensional (2D), three dimensional (3D), or motion and time-based. In Visual Arts, the students were required to submit a finished art product, as opposed to the task in Design course, which was in the form of evidence of process. The artist statement that accompanied the resolved artwork could describe the production process, however this artist statement did not contribute to the student's score directly.

Upon the WACE practical examination submission, the participating students' artworks and the documents they provided were digitised. In this digitisation process, researchers from the main project took digital photographs and videos of the artworks at the examination site, where the artworks were already prepared for WACE marking. The original photographs and videos were later processed to fit the technical requirement of the assessment software and Internet bandwidth limitation. Details of this process were discussed in Chapter 3. Digital representations taken for the student artworks comprised photographs of fully installed artwork, close-ups, and videos. Three-dimensional artworks that could fit on a revolving table were also video recorded in virtual reality (VR) video format.

In Visual Arts, the widely varied types of artwork and the quality of the digital representations were potentially the main source of concern for the scoring processes. The artworks submitted for the WACE examination could be in the form of drawing, sculpture, a collection of different works, textiles, and many more. There were variations in physical aspects such as size, medium, and dimension; as well as in intellectual aspects such as creativity, innovation, ideas, and style.

In the Analytical marking process, these variations could result in different judgements among assessors in the kind of qualities that constituted a score. In the Comparative Pairs judgements the problem might arise from the difficulty in comparing two very different types of artworks and deciding on the better. Besides making scoring problematic, the wide range of artwork types also made the digitisation challenging. As a result, in several cases the digital representations of the artworks provided the assessors with constrained information on the artworks, and this could create a discrepancy between scores obtained from the digital scoring processes and the WACE scores in which the original artworks were assessed. Examples of notes from the ACJ system indicating these artworks were: *I am making the judgement without being able to see the close ups on B but the composition is less appealing* and *message not well conveyed without the support of the artist statement*. Problems associated with the nature of the student work could consequently affect the reliability of the scoring results.

Constraints from the digitisation process

As for the digitisation of the Design portfolios, there were also several problems that might affect the quality of the digital representation of the Visual Arts works. A guideline to accommodate each type of work in the digitisation of student work was constructed with consultation from experienced Visual Arts markers and the Curriculum Council. Despite the researchers' best efforts to follow the guideline as closely as possible, there were still several problems that may have compromised the quality of the digitised version of student work. The main problems were associated with time constraints and the installation of the artworks. These problems are now each discussed.

Time constraints

The digitisation process for the online scoring had to be completed in one day, which was the day between the submission date and the start date for the WACE marking. For the WACE submission, all Visual Arts students' works were submitted to a location where they were catalogued and arranged for marking.

The research team was divided into several groups with each group assigned a list of students. Each team then searched for each participating student work, arranged the work by following the installation photograph provided by the student, and proceeded to take the still pictures and videos as carefully as possible to prevent any disturbance on the artworks.

The digitisation process was time consuming while the allocated time was very limited. Time limitation made it impossible to apply sufficient digitisation methods such as checking the quality of the photographs and videos on site, as well as to use extra equipment such lightings and backdrops. Therefore despite every care taken to ensure the best quality for the digital representation, there was a possibility that some digital representations were not as representative as they needed to be.

Artwork installation

Beside the limited time available in the digitisation process, problems also arose from the installation of the artworks. All artworks were delivered for submission to the Curriculum Council collectively by the schools, therefore when they arrived in the WACE marking site they were in their delivery packaging and were not installed. Artworks that consisted of separate parts could be difficult to assemble, especially when utmost care should be employed to prevent damage to the work. Each artwork that was about to be digitised was arranged to match as closely as possible the installation picture that the student submitted with it, however for several artworks this could be challenging. An example would be a two-piece gown made of colouring pencils that was supposed to be fully arranged on a mannequin but was submitted with only one piece, the torso, installed while the skirt was not. For artworks that were difficult to install, the digital representations could not perfectly represent the artworks as intended by the students who created them, and thus, the scores may not accurately represent the quality of their work.

Technical limitations

For the digitisation process the research team was assisted by two professional photographers. The team used SLR digital cameras and HD video recorders to aim to create the highest quality digital representation of the student work that was possible under the circumstances. However, there were two major technical problems that the team encountered. In Visual Arts, the artworks could be created using various media. Several of the artworks that needed to be digitised were created in media that were difficult to be digitally captured, for example glass and/or Perspex. Reflection, change of colour, missing details and patterns, and the lack of a three-dimensional perspective were among the technical problems that arose from the digitisation process. Combined with the time limitation, this problem was likely to result in disparity between the real artwork and the digital representation when judging for assessment purposes.

Beside technical problems during the digitisation process, there was also a problem with the file size. For the Analytical marking and the Comparative Pairs judgements processes, the assessors scored the digital version of student work online. In order for this online scoring to be smooth, the file size for the pictures and videos had to be as small as possible without sacrificing the quality of the pictures and videos. A guideline on file format and size was created, as was discussed in Chapter 3, to ensure that the quality of the digital representation was optimal. Nevertheless, this could still be a possible source of disparity between the real artwork and the digital representation, which could affect the reliability of the results from the online scoring processes.

Analysis of the Scoring Data

This section presents the results from an analysis on the scoring data from the three sources, which were the Analytical marking of the digital representations, the official WACE analytical marking, and the Comparative Pairs judgements. Several Visual Arts course educators from secondary and tertiary levels were invited to be assessors for both the Analytical marking and the Comparative Pairs judging, and fifteen decided to be involved. Three of these assessors, who were experienced WACE markers, were also the markers in the Analytical marking process.

Analytical marking for Visual Arts course

For the Visual Arts course three assessors marked the student artwork analytically by using the rubric in the online tool; the same rubric that was used for the WACE practical examination score from the original artworks (marked by others). The rubric consisted of five criteria with maximum score points ranged from 5 to 12 for criteria, with a total score of 40. The complete rubric that was used in these two scoring processes can be viewed in Appendix D. The criteria were titled:

C1: Creativity and Innovation (0 – 6)

C2: Communication of Ideas (0 – 5)

C3: Use of Visual Language (0 – 12)

C4: Use of Media and/or Materials (0 – 5)

C5: Use of Skills and/or Processes (0 – 12)

Processes and time taken for marking

The Filemaker Pro scoring interface was equipped with a timer to record the time spent by the assessors to assess the PDF file, the video, and for three-dimensional work, the 360° Virtual Reality (VR) video. This timer started at the beginning of each scoring session and recorded the total time each assessor needed to assess an artwork. The recorded time could include unintended breaks during the scoring as well as the time needed for the digital files to load. The size of the digital files was quite big and therefore it could take some time to load.

The average time taken for the Analytical marking was 9.9 minutes per artwork, ranging from about five to twenty minutes for each artwork. The total amount of time for marking the 75 artworks by the three assessors was 37 hours. It was noted that there were files that were bigger than the others hence took more time to load. In comparison for the Design course, the file size for every portfolio was relatively similar, therefore long marking time could indicate difficulty in marking. In Visual Arts, however, the file size as well as the number of files varied, therefore the marking time recorded could be affected by the loading time and could not be used to indicate difficulty in marking.

Scores from marking

The Filemaker Pro database scoring interface provided the marking rubric for Visual Arts on which the assessors assigned a score to each criterion for each artwork. Results from this marking process were recorded in the database then imported to a spreadsheet and analysed by using SPSS. For each student a score for each criterion was recorded and then summed to generate a total score. The structure of these scores from the Analytical marking is as shown in Table 5.1. This structure was designed to make analysing the scoring results based on criteria, assessors, or schools easier. Analysis of scores from Analytical marking was based on schools, assessors, and comparison with WACE official marking result.

Each school participating in the study was assigned an identification code that consisted of two letters, the first letter was V, for Visual Arts, the second letter was the school code. Each student participating in the study was assigned an identification code that consisted of two letters from the school code followed by three digits of number. The structure of the Analytical marking scores is as shown in Table 5.1The purpose of this coding was to maintain the privacy of both the schools and the students involved. Only the researchers involved in the project had access to this coding. There were 75 students from ten schools involved in this project, with the number of students varying from only three in schools VC and VH to thirteen in school VS. The scoring processes and results are discussed in the next sections.

Table 5.1
Structure of Analytical Marking Scores

ID	Criterion (Max Score)					Total
	C1 (6)	C2 (5)	C3 (12)	C4 (5)	C5 (12)	(40)
VC901	3.3	2.7	6.7	3.0	6.7	22.3
VC902	5.0	3.0	6.7	3.3	7.7	25.7
VC903	2.0	1.7	3.0	2.0	3.7	12.3

Analysis of scores based on schools

Table 5.2 presents the Analytical marking score data, based on schools. This analysis was intended to examine possible patterns or peculiarities among schools for each criterion. In this analysis, the means and standard deviations for each school for each criterion were calculated to be compared.

Table 5.2

The Mean Score for Each School per Criterion from the Analytical Marking Process

School	N	Score (SD)					
		C1 (6)	C2 (5)	C3 (12)	C4 (5)	C5 (12)	Total (40)
VC	3	3.4 (1.5)	2.4 (0.7)	5.4 (2.1)	2.8 (0.7)	6.0 (2.1)	20.1 (6.9)
VH	3	4.9 (0.8)	3.8 (0.8)	8.0 (1.9)	3.8 (0.5)	8.4 (1.9)	28.9 (5.9)
VJ	10	3.7 (0.7)	2.8 (0.6)	6.5 (1.0)	3.0 (0.4)	7.0 (1.0)	22.9 (3.4)
VK	9	4.3 (0.9)	3.3 (0.7)	7.8 (1.8)	3.6 (0.7)	8.3 (1.7)	27.3 (5.8)
VL	11	2.9 (0.4)	2.3 (0.4)	5.0 (0.9)	2.4 (0.3)	5.2 (0.8)	17.7 (2.7)
VN	10	4.3 (1.3)	3.3 (0.9)	7.6 (2.1)	3.5 (0.8)	7.9 (2.0)	26.7 (7.0)
VO	4	3.8 (0.8)	2.9 (0.7)	6.2 (1.3)	2.8 (0.6)	6.8 (1.5)	22.6 (4.9)
VP	7	4.5 (0.4)	3.4 (0.4)	7.7 (1.1)	3.3 (0.2)	7.9 (1.1)	26.8 (2.9)
VQ	5	3.5 (0.9)	2.5 (0.7)	5.7 (1.6)	2.9 (0.5)	6.6 (1.2)	21.2 (4.7)
VS	13	3.5 (0.6)	2.7 (0.5)	6.0 (1.1)	2.8 (0.4)	6.2 (0.9)	21.2 (3.2)
MEAN	75	3.8 (0.9)	2.9 (0.7)	6.6 (1.7)	3.0 (0.6)	6.9 (1.7)	23.2 (5.5)

Table 5.3 shows the same scores in percentages. This conversion helped in making differences among schools or criteria more noticeable. There was no school with a criterion that had a total mean score particularly different to another. The mean scores for all criteria for all schools were within one standard deviation of the total mean scores. This indicates that while there was variation in the way each criterion contributed to the total school score, there was no particular criterion that contributed substantially more than the others to the total school score.

Table 5.3
The Mean Score in Percentage for Each School per Criterion from the Analytical Marking Process

Case	N	Score (SD)					
		C1 (6)	C2 (5)	C3 (12)	C4 (5)	C5 (12)	Mean (40)
VC	3	57.4	48.9	45.4	55.6	50.0	50.3 (17.3)
VH	3	81.5	75.6	66.7	75.6	70.4	72.2 (14.7)
VJ	10	61.7	55.3	53.9	59.3	58.1	57.2 (8.5)
VK	9	72.2	66.7	65.1	71.1	68.8	68.2 (14.4)
VL	11	48.0	45.5	41.4	47.9	43.4	44.3 (6.8)
VN	10	71.7	66.7	63.6	70.0	66.1	66.8 (17.4)
VO	4	62.5	58.3	52.1	56.7	56.9	56.5 (12.1)
VP	7	74.6	68.6	64.3	65.7	65.5	66.9 (7.3)
VQ	5	58.9	49.3	47.8	57.3	55.0	53.0 (11.8)
VS	13	58.1	53.8	49.8	56.9	51.5	52.9 (8.1)
MEAN	75	63.3 (15.8)	58.0 (14.7)	54.6 (14.2)	60.9 (12.8)	57.7 (13.9)	58.1 (13.9)

The overall mean score for each criterion ranged between 54.6% for criterion C3 – *Use of Visual Language,* to 63.3% for C1- *Creativity and Innovation*. The overall mean score across all criteria was 58.1% (SD=13.9%) and there was no criterion with a total mean score more than one standard deviation difference to that overall mean score. This further indicates that in general there was no criterion that was scored substantially differently to the rest.

However, there was a distinct pattern in the mean score percentage for each school in each criterion, as could be observed in *Figure 5.1*. For almost every school, the criteria with the highest mean score percentage to the lowest were C1, C4, C2, C5 and C3. The only exception was school VO, which had a slightly higher mean score percentage for C5 than for C4. This general pattern among schools indicated a possibility of inequality among criteria with several criteria required qualities that were either more difficult to attain or more difficult to demonstrate than the others. It also could indicate a gap between the assessors' expectation and the students' understanding on how several criteria were realised in the artworks. This is despite the lack of large differences between the mean scores in each criterion and the overall mean scores mentioned previously. Even though the differences might be statistically insignificant, they were consistent across schools regardless of the number of participating students in each school.

Overall there was consistency in the position of each school compared to the others. Schools VH, VP, VK, and VN were consistently above the total mean for all criteria. School VH had higher mean scores than the other schools in all criteria. Schools VP, VK, and VN had similar mean scores except in C4 with school VP having a lower mean score compared to schools VK and VN. Schools VJ and VO had mean scores that were quite similar to the average mean scores. Schools VS, VQ, and VC had relatively similar mean scores across criteria while school VL consistently had the lowest mean scores. This pattern could indicate agreement among criteria, which contributes to the construct validity of the assessment. However, further analyses needed to substantiate this claim were not conducted because it was not relevant to the aim of this study, which was focussed on the Comparative Pairs judgements.

This pattern could also indicate the influence of school culture; such as collective academic characteristics of Visual Arts students in each school (e.g. persistence, understanding, intelligence), specific teaching methods, teaching-to-the-test approach, availability of school facilities, and others; on student achievement in practical assessment that could be an interesting and important topic for a further study.

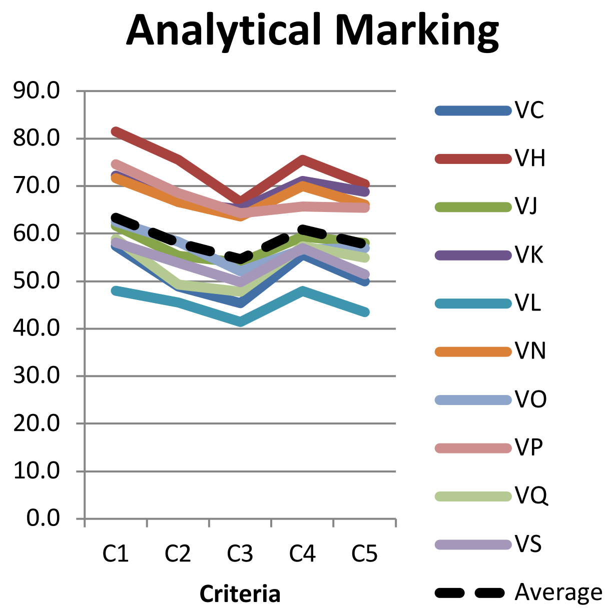


Figure 5.1 Analytical marking result for each school per criterion.

Analysis of scores based on assessors

A summary of the scores obtained from each assessor for each criterion is presented in more detail in Table 5.4. In general there was agreement between the three assessors. However, it appears that Assessor 2 tended to be more generous in scoring, which could be seen from the underutilisation of the lower scores and the relatively higher average score per criterion compared to the other assessors. Most of the standard deviations for the scores given by Assessor 2 were also lower than the other assessors, showing that the spread of the scores was closer.

Table 5.4

Descriptive Statistics on Marking for All Students by Each Assessor

Set of Criteria	Assessor	Possible	Range	Mean	Std. Deviation	Mean (%)*
Creativity and Innovation	1	6	1-6	3.5	1.3	58.2
	2	6	2-6	4.1	1.0	68.0
	3	6	1-6	3.8	1.2	63.8
	Average	6	2-6	3.8	0.9	63.3
Communication of ideas	1	5	1-5	2.7	1.0	53.3
	2	5	2-5	3.2	0.8	63.7
	3	5	1-5	2.9	1.0	57.1
	Average	5	1.7-4.7	2.9	0.7	58.0
Use of visual language	1	12	1-11	6.0	2.2	50.1
	2	12	2-11	6.9	1.8	57.8
	3	12	2-11	6.7	2.1	56.0
	Average	12	3-11	6.6	1.7	54.6
Use of media and/or materials	1	5	1-5	2.9	0.8	57.1
	2	5	2-5	3.2	0.7	64.3
	3	5	1-5	3.1	0.8	61.3
	Average	5	2-5	3.0	0.6	60.9
Use of skills and/or processes	1	12	2-11	6.1	2.3	51.1
	2	12	4-11	7.3	1.8	61.1
	3	12	3-11	7.3	2.1	61.0
	Average	12	3.7-11	6.9	1.7	57.7

*Percentage of the mean average

A calculation on the differences between the scores given by the three assessors showed that the largest score difference was 21 (out of 40) with a mean of 8.5 (SD=4.3). Correlations between assessors' scores were significant but relatively low with correlation coefficients ranging between 0.51 and 0.56 ($p<0.01$), indicating only moderate agreement between assessors even though they were all experienced Visual Arts educators using the same Analytical marking criteria. This relatively low level of agreement indicates a low level of inter-rater reliability in the scores resulting from the Analytical marking process.

Table 5.5 shows the correlation coefficients between assessors while *Figure 5.2* depicts the scatter plots of the scores given by the Analytical assessors.

Table 5.5
Correlations Between Scores from the Three Assessors

(N=75)	Assessor1	Assessor2	Assessor3
Assessor1	1	0.54**	0.51**
Assessor2		1	0.56**
Assessor3			1

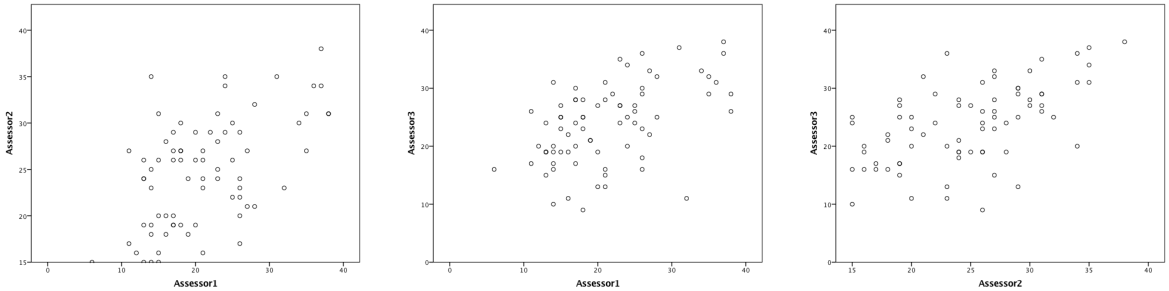


Figure 5.2 Scatter plots between scoring results from the Analytical marking assessors.

Despite the correlation between scoring results from the three assessors being low, there were not many artworks with large differences between the scores from each assessor. There were only three out of 75 artworks (4%) with a difference of more than 2 standard deviations between the scores from each assessor.

Comparison between analytical and WACE practical marking scores

The WACE practical scores for participants in the study were provided by the Curriculum Council authority. These scores were generated from assessors who marked the students' original artworks using the same rubric to that used in the study for the Analytical marking of the digital version of the students' artworks. That is, the difference between these two scoring methods was only the form of the artworks being marked. In the official WACE marking the assessors marked the original artworks while in the Analytical marking, the assessors marked the digitised version.

As for the Analytical marking, in the WACE marking there were several assessors with each artwork being marked by at least two assessors. In case of extreme dissimilarities in marking, a meeting was held to discuss the differences and to obtain an agreed score. The WACE practical score used in this study was the mean of the scores from the assessors, or the score from the reconciliation meeting. A summary of the results from the Analytical marking and WACE scores is shown in Table 5.6.

Table 5.6
Descriptive Statistics on Analytical Marking and WACE

		N	Range	Mean	Std. Deviation
Analytical Marking	Assessor1	75	6.0 – 38.0	21.2	7.4
	Assessor2	75	15.0 – 38.0	24.8	5.8
	Assessor3	75	9.0 – 38.0	23.8	6.9
	Average	75	12.3 – 37.7	23.9	6.9
	Average (%)	75	30.8 – 94.2	58.1	13.9
WACE	WACE Practical	75	10.0 – 40.0	25.3	6.3

Compared to the result from the WACE practical scores, the mean from the Analytical marking was slightly lower (23.9 cf. 25.3). The three Analytical marking assessors utilised different minimum scores, 6.0, 9.0 and 15.0, with the same maximum score of 38.0, while the WACE scores ranged between 10.0 and 40.0. The means of these scores were only slightly different to the WACE mean score. The score distributions varied considerably with Assessor 2 having the least spread scores with the standard deviation of 5.8 and Assessor 1 having the widest at 7.4.

Comparative Pairs judgements for Visual Arts course

Data from the Comparative Pairs judgements were obtained from the ACJ system using the judgements done by 14 Visual Arts assessors. All assessors were qualified teachers or academics in Visual Arts education. Before the scoring process commenced the researchers hosted a four-hour workshop with the Visual Arts assessors. This workshop had two main purposes; the first was to decide on a holistic criterion upon which the Comparative Pairs judging was to be based. This criterion was based on the marking rubric developed for the official WACE practical examination. In this workshop, the assessors discussed and decided on a holistic criterion for the Comparative Pairs judgement which was:

Holistic Criterion: Judgement about performance addresses students' ability to creatively use visual language, materials and processes to skilfully communicate an innovative idea in a resolved artwork.

The WACE marking criteria upon which this holistic criterion was based were:

- *Creativity and innovation* - Artwork/s is outstanding, showing exceptional creativity and innovation and the emergence of a distinctive style.
- *Communication of ideas* - Ideas are skilfully realised and powerfully communicated in sophisticated and highly coherent resolved artwork/s.
- *Use of visual language* - Extensive and sophisticated application of visual language in the artwork/s. Complex and highly resolved visual relationships are evident.
- *Use of media and/or materials* - Highly discerning selection and refined use of media and/or materials demonstrating sensitive application and handling.
- *Use of skills and/or processes* - Extensive and sophisticated selection and application of skills and processes.

These criteria were also the criteria that were used in the Analytical marking process for the digital artworks.

The second purpose of the workshop was to introduce to the assessors the judging interface of the ACJ online system and to ensure that there was a common understanding on how to use the holistic criterion. At the end of the workshop the assessors started judging the first few pairs in the first judging round. The remainder of the judging process was conducted off-site at home or workplace.

ACJ System Data on Comparative Pairs Judging

The Comparative Pairs judgements data were obtained from the ACJ system. The system created the pairings from which the assessors judged the better one in each, and subsequently ranked the students based on those judgements. At the end of the whole judgement process, which consisted of many rounds, the system ranked the artworks on a parameter measurement scale in Rasch logits, and provided information on judgements sessions as well as an analysis of reliability and individual artwork or assessor misfits. Features from the ACJ system that were used in this study have been discussed more fully in Chapter 3.

In the first rounds, the ACJ system paired the artworks randomly then more adaptively, resulting in gradually faster judgements and more accurate scoring results. *Figure 5.3* shows how the standard error bars of the parameter value improved between the first and the last round. The graph curve also became smoother, which indicated that the rank of the student was getting more closely together and the difference in quality became finer. However, there was a notable difference between the Visual Arts and Design graphs in terms of the location distribution and standard errors. These differences are discussed in Chapter 6.

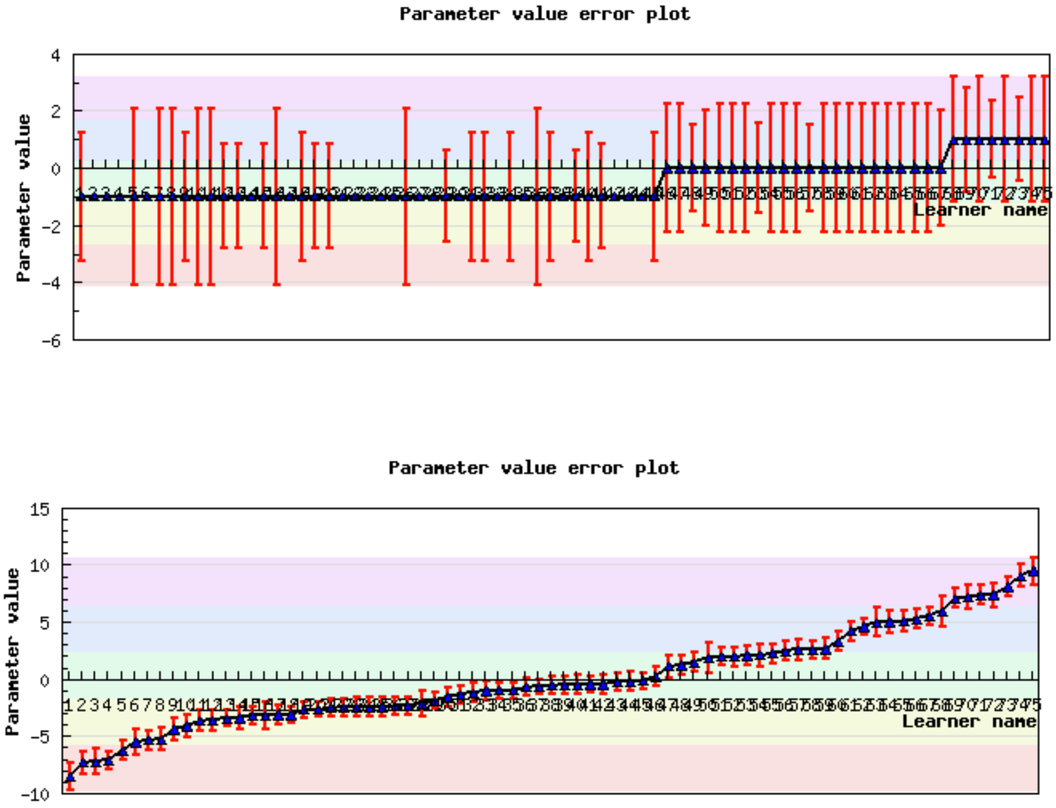


Figure 5.3 Parameter value error plot from the first and last rounds.

By the end of the thirteenth round, the reliability coefficient reached .959, and the judgement process was concluded because it was understood that after this point it was likely that there would be little increase in the reliability coefficient. This high reliability level represented both the internal reliability and the inter-rater reliability of judgement among judges (Kimbell, 2007).

Table 5.7 shows how the reliability coefficients increased for every round of judgement. Related to the discussion of the ACJ system in Chapter 3, the first six rounds had not resulted in a meaningful reliability coefficient. Table 5.7 shows the increase in the reliability coefficients. From the seventh round onward there was a steady increase in the reliability coefficient as more fine-tuning in the pairing was created and artworks of more similar quality were paired to be judged.

Table 5.7
Reliability Coefficients From the Last Eight Rounds of Comparative Pairs Judgements

Round	r
6	*
7	*
8	0.900
9	0.930
10	*
11	0.950
12	0.956
13	0.959

* These values were not recorded at the time

Consistency of the Assessors and Judgements

During the judgement process, the ACJ system compared each judgement made by the assessors with the overall judgements. This process provided the researchers with information on the consistency of the assessors in misfit statistics data. These misfit data included the mean residual, the weighted mean square, and the unweighted mean square. The consistency statistics from the ACJS is as shown on Table 5.8.

The mean residual for all assessors were around the mean of 0.46, which indicates that all assessors were relatively in agreement with one another. The misfit statistics shown by the weighted mean square had a mean of 1.37 (SD=0.30) with only four assessors (Assessors 2, 5, 8 and 14) having a mean difference that was more than one standard deviation but less than two standard deviations. There was no sufficient data from Assessor 1, therefore this assessor was excluded from the analysis. Among all 497 judgements there were only 42 (8.5%) judgements that the system identified to be inconsistent.

Table 5.8
Consistency Statistics for Assessors for the Visual Arts works

Assessor	Count	Mean Residual	Unweighted mean square	Unweighted Z	Weighted mean square	Weighted Z
1	1	0.50	1.00	0.00	1.00	0.00
2	47	0.52	5.97	6.62	1.90	5.79
3	32	0.40	11.55	5.91	1.06	0.49
4	15	0.48	1.08	0.58	1.07	0.56
5	50	0.43	31.96	11.84	1.73	4.57
6	32	0.44	6.78	4.53	1.45	3.18
7	42	0.41	1.04	7.23	1.13	1.18
8	32	0.52	12.59	2.97	1.69	4.63
9	39	0.45	1.27	0.59	1.07	0.92
10	40	0.49	20.89	2.88	1.67	4.98
11	32	0.47	11.97	6.47	1.49	2.83
12	33	0.39	9.62	4.95	1.35	1.88
13	32	0.52	5.98	6.11	1.51	3.46
14	30	0.45	0.90	-1.02	0.92	-0.97
15	40	0.47	2.13	1.85	1.46	2.84
	Mean:	0.46	8.31	4.10	1.37	2.42
	S.D.:	0.04	8.47	3.32	0.30	1.97

Processes and Time Taken for Judging

There were 15 assessors involved in the Comparative Pairs judgements process, however there was not enough activity from one of them, therefore only results from the other 14 judges were analysed. There were 497 judgements in almost 45 hours made in total, averaging at 5:24 minutes per judgement. Each judgement took from 2.22 to 9.18 minutes per judgement, with fluctuating average time. It should be noted that this amount of time could include breaks that might be taken by the assessors during judgement sessions. However, the system calculations tried to make allowances for extreme values. Table 5.9 shows the estimated time for each round in the Comparative Pairs judgements process.

Table 5.9
Estimates of Time Taken Making Judgements for Comparative Pairs Judging of the Visual Arts Works

Round	Total time (hrs)	Judgements	Average Time per Judgement (hrs)
1	5:03:23	38	0:07:59
2	1:31:14	25	0:03:38
3	1:53:53	26	0:04:22
4	1:11:41	26	0:02:45
5	1:05:24	25	0:02:36
6	2:30:25	35	0:04:17
7	3:02:55	36	0:05:04
8	2:24:02	36	0:04:00
9	3:46:57	37	0:06:08
10	2:23:11	37	0:03:52
11	3:15:30	37	0:05:17
12	2:02:43	34	0:03:36
13	3:12:48	40	0:04:49

Scores from Comparative Pairs Judgements

Scores from the Comparative Pairs judgements were obtained from the ACJ system. At the end of the judgement session the ACJ system provided a summary of the location parameter for each student, including the inconsistency statistics. The structure of this summary was as displayed in Table 5.10.

Table 5.10

Sample of Student Location Parameter Result from the ACJ System

Student ID	Parameter	SE	Unweighted mean square	Unweighted Z	Weighted mean square	Weighted Z
VC901	-1.19907	0.88	10.35	2.17	2.11	3.53
VC902	-4.40156	0.94	1.40	1.23	1.17	1.16
VC903	-8.51322	1.15	9.63	23.51	1.48	1.50
VH901	-3.40199	0.95	51.07	41.18	1.73	2.92
VH902	5.96092	1.32	1.23	0.87	1.24	0.95
VH903	4.57262	0.76	7.74	21.50	2.04	3.66
VJ901	1.94057	0.85	28.05	4.69	1.88	3.80
VJ902	0.198931	0.82	4.17	3.12	1.76	2.38
VJ903	2.06049	0.80	2.81	3.29	1.71	2.69
VJ904	4.21855	0.84	54.38	5.41	1.06	0.36
Mean:	0.00000	0.81	9.39	4.81	1.42	1.86
S.D.:	4.12424	0.21	16.07	7.51	0.36	1.43

This judgement and analysis process resulted in a score set that ranged from -8.513 to 9.483. The frequency distribution of the location parameter followed a normal distribution with an average of 0.0000003, which was very close to 0 as expected in a Rasch modelling distribution. The graph of the frequency distribution is displayed in *Figure 5.4*. This location parameter was based on the Rasch dichotomous model that was employed by the ACJ system as discussed in Chapter 3. From the 75 artworks assessed in this scoring method, only four artworks (5%) had a weighted mean square value above 2 SD from the average mean square value. This suggested that the judgements were less conclusive on these four artworks than the rest.

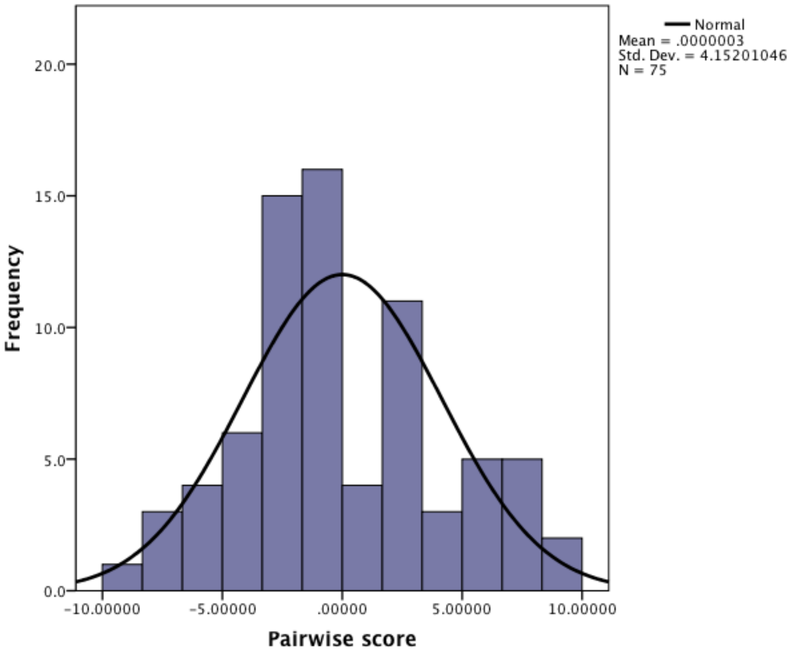


Figure 5.4 Frequency distribution of Comparative Pairs scores

The ACJ system judged that, assuming the scores represented a population of about 6 SD's wide and that bands 3 SE's apart are distinguishable, there were up to 10.2 reliably distinct bands. These bands could be used for grading the artworks but because they were not pertinent in this study, they are not discussed. A normality test further showed that the parameter distribution followed a normal distribution, as indicated in Table 5.11, and explained below the table.

Table 5.11
Normality Tests Results

Descriptives

			Statistic	Std. Error
Pairwise score	Mean		0.0000003	0.47943287
	95% Confidence Interval for Mean	Lower Bound	-0.9552905	
		Upper Bound	0.9552912	
	5% Trimmed Mean		-0.0560468	
	Median		-0.5424870	
	Variance		17.239	
	Std. Deviation		4.15201046	
	Minimum		-8.51322	
	Maximum		9.48312	
	Range		17.99634	
	Interquartile Range		5.23304	
	Skewness		0.347	0.277
	Kurtosis		0-.392	0.548

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Pairwise score	0.114	75	0.016	0.974	75	0.130

a. Lilliefors Significance Correction

The z-score for skewness for this parameter is:

$$z = \frac{Skewness}{Std.Error} = \frac{0.347}{0.277} = 1.253$$

This skewness z-score is within the range of ± 1.96 (p<.05), indicating that the distribution could be considered symmetrical and not significantly skewed. The z-score for kurtosis is:

$$z = \frac{Kurtosis}{Std.Error} = \frac{-0.392}{0.548} = -0.715$$

This kurtosis z-score is also within the range of ± 1.96 (p<0.05), indicating that the distribution is mesokurtic, or follows a normal distribution. The significant probability value of the Shapiro-Wilk statistic is 0.13, which is larger than α= 0.05, suggesting that the parameter distribution is normally distributed.

Summary of scoring results

Shown in Table 5.12 is a summary of the scores and ranks from all three methods of scoring for each school. The location parameter (logits) resulted from the Comparative Pairs judgement was not on the same scale as the other two scoring methods, therefore the result was rescaled using the mean and standard deviation from the Analytical marking result. The Analytical marking result was used as a baseline because the Comparative Pairs judgements method used a holistic criterion that was based on the criteria for the Analytical marking, and also because unlike the WACE marking process, both the Analytical marking and the Comparative Pairs judgements processes used the same digital representations of student artwork. The Comparative Pairs judgement scores were in logits as the ACJ system used the Rasch dichotomous model to calculate the student ability and criterion difficulty. These scores were rescaled and ranked to parallel the results from the other scoring processes to better describe and compare the results.

The school mean scores in each scoring method were mostly around the total mean score. The Analytical marking mean score for each school was within one standard deviation difference from the total mean score from this method. In the Comparative Pairs judgements, school VC was the only one that was lower than the total mean score by slightly more than one standard deviation. In the WACE marking, school VL was lower than the total mean score by slightly more than one standard deviation while school VH was quite far higher than the total mean score.

Table 5.12
Scoring Result Summary from All Scoring Methods for Each School

	N	Assessors Average		CP		WACE Practical	
		Score (40)	Rank	Logits (Rescaled)	Rank	Score (40)	Rank
VC	3	20.1 (6.9)	45.8 (26.4)	16.9 (4.9)	62.0 (16.1)	20.5 (9.7)	46.8 (26.5)
VH	3	28.9 (5.9)	17.2 (18.1)	26.4 (6.7)	28.0 (29.6)	34.7 (7.6)	12.5 (18.2)
VJ	10	22.9 (3.4)	37.7 (16.9)	23.9 (3.4)	33.4 (16.5)	26.2 (4.7)	34.7 (19.3)
VK	9	27.3 (5.8)	23.3 (19.9)	27.4 (5.8)	23.9 (20.8)	30.2 (5.7)	20.3 (18.8)
VL	11	17.7 (2.7)	62.3 (12.8)	19.7 (5.2)	51.3 (21.7)	18.3 (4.3)	62.6 (13.5)
VN	10	26.7 (7.0)	27.6 (24.0)	25.1 (7.4)	33.3 (25.9)	29.6 (5.6)	23.8 (19.4)
VO	4	22.6 (4.9)	38.1 (23.1)	21.7 (1.3)	41.8 (9.3)	22.9 (4.9)	47.5 (20.6)
VP	7	26.8 (2.9)	20.7 (10.0)	24.4 (2.7)	29.9 (13.0)	27.3 (3.3)	28.6 (14.0)
VQ	5	21.2 (4.7)	47.2 (21.7)	24.0 (4.3)	33.4 (19.5)	24.9 (2.4)	39.0 (11.1)
VS	13	21.2 (3.2)	44.6 (13.9)	21.6 (5.4)	45.5 (22.1)	22.6 (3.7)	48.5 (14.3)
ALL	75	23.2 (5.5)		23.2 (5.5)		25.3 (6.3)	

The *Rank* columns show the mean for the ranks of the artworks in each school. The ranking of schools from the three scoring methods were quite different to the scores, partly because of the small sample size in each school. Because the number of students in each school was small, a small discrepancy between scores could create a larger discrepancy in overall rankings. *Figure 5.5* further illustrates the variation of the overall score means obtained from the three scoring processes: Analytical (average of the two assessors for each student), Comparative Pairs, and WACE practical scores.

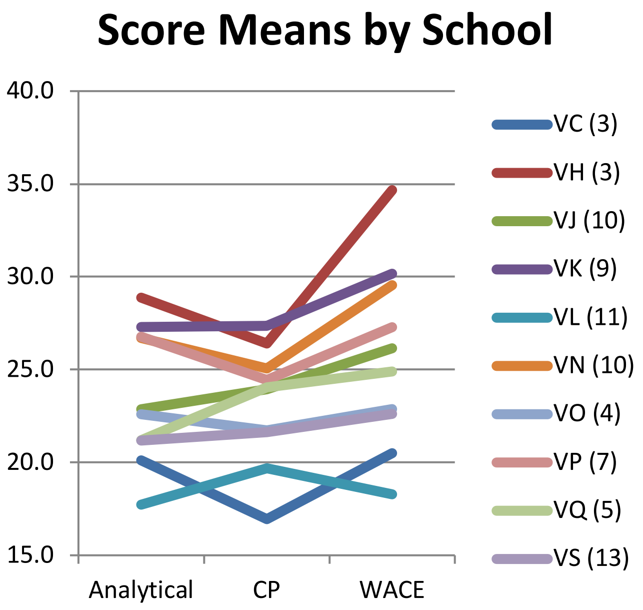


Figure 5.5 Score means by school from each scoring method.

In general, the school mean scores in the three scoring methods varied with the mean scores from the Comparative Pairs judgements seemed to be the most different while the mean scores from the other two methods were mostly similar. The mean scores from the Comparative Pairs judgements for schools VH, VN, VP, and VC were lower compared with the other two methods. VL was the only school in which the Comparative Pairs judgements gave a higher mean score than the other methods. In schools VJ, VQ, VO, and VS the mean scores from all methods were relatively similar. School VH had a WACE mean score that was well above other schools across methods.

Figure 5.6 shows how the rank of each school from the different methods varied from the mean across the schools. Only three schools; VK, VJ, and VS were ranked relatively similarly across methods while the rest of the schools had quite different ranks. Rankings from the Analytical marking and the WACE marking were quite similar for schools VC, VL, and VN.

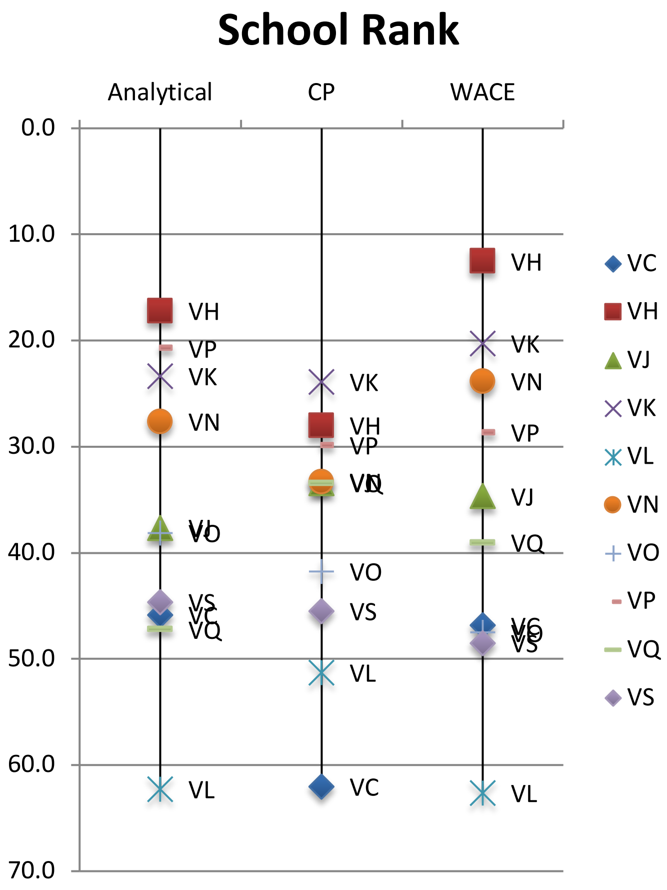


Figure 5.6 Rank means by school from each scoring method.

This presentation of data serves as a preliminary analysis that was aimed to observe possible patterns that might emerge when the schools were compared. Unlike in the Design course, in the Visual Arts course the indication that there could be typical academic characteristics of Visual Arts students in each school that might influence student achievement was less prominent. More detailed analyses are discussed in the following sections.

Comparison between scores from three sources

A correlation analysis was conducted between the scores and rankings that resulted from the three scores. This analysis was done to examine the similarity of the scoring results as part of the validity analysis. The correlation analysis result is shown in Table 5.13.

In general, the correlations between the scores from the three methods of scoring were significant and high, with the correlation coefficient between the Analytical and WACE markings being the highest (r=0.84, p<0.01), followed by the correlation between Analytical marking and Comparative Pairs judgements (r=0.79, p<0.01), and between WACE marking and Comparative Pairs judging (r=0.74, p<0.01). The correlations between rankings were relatively similar to the correlations between sets of scores.

Table 5.13
Correlations Coefficients Between Scores from the Three Methods of Scoring

(N=75)	Assessor1	Assessor2	Assessor3	Analytical	CP	WACE Practical
Assessor1	1	0.54**	0.51**	0.84**	0.68**	0.70**
Assessor2		1	0.56**	0.82**	0.72**	0.75**
Assessor3			1	0.83**	0.58**	0.71**
Analytical				1	0.79**	0.86**
CP					1	0.74**
WACE Practical						1

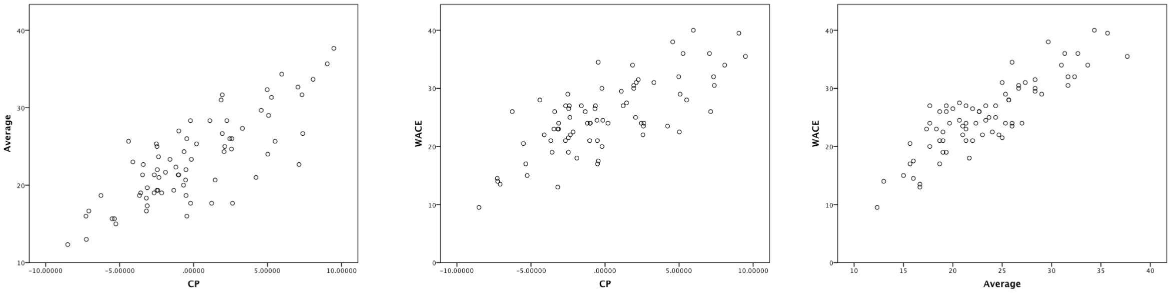
**. Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the 0.05 level (2-tailed).

Table 5.14
Correlations Coefficients Between Ranks from the Three Methods of Scoring

Rank of	Assessor1	Assessor2	Assessor3	Analytical	CP	WACE Practical
Assessor1	1	0.49**	0.51**	0.80**	0.62**	0.64**
Assessor2		1	0.56**	0.81**	0.68**	0.72**
Assessor3			1	0.85**	0.56**	0.70**
Analytical				1	0.73**	0.82**
CP					1	0.67**
WACE Practical						1

**. Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the 0.05 level (2-tailed).

The high correlations between scores obtained from the three scoring methods indicated similarities in the results regardless of the different processes and scoring media (i.e., digital or physical). The scatter plots for these relationships are as shown in *Figure 5.7*.



Note: Average = The mean scores from the two Analytical marking assessors

Figure 5.7 Scatter plots between scoring results from the three methods.

Even though the correlations between scores from the Analytical marking and the other two methods were significantly high, the correlations between scores from individual assessors were low, as was discussed in the Analytical marking for Visual Arts course section. Considering the *Analytical* scores were the mean scores from the three assessors, this suggested that despite the low level of agreement among Analytical marking assessors, the differences were in some sense cancelled out in the averaging process. *Figure 5.8* and *Figure 5.9* depict the scatter plots for scores between each assessor and other scoring methods.

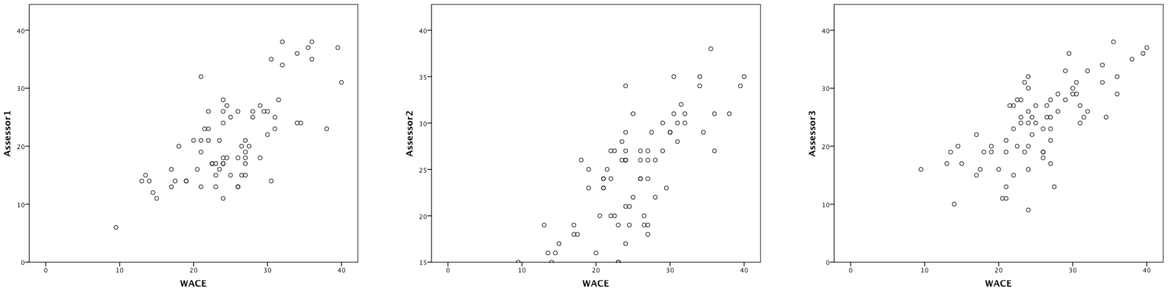


Figure 5.8 Scatter plots between scores from each assessor and WACE.

The correlations between the scores for each assessor and the WACE scores were moderate and quite similar for all three assessors, with correlation coefficients ranging between 0.70 and 0.75 ($p<0.01$). When compared with results from the Comparative Pairs judgement, the correlation coefficients were more varied, with Assessor 2 having the highest coefficient of 0.72 (p<0.01), followed by Assessor 1 (r= 0.68, p<0.01) and Assessor 3 with a low correlation of 0.58 (p<0.01).

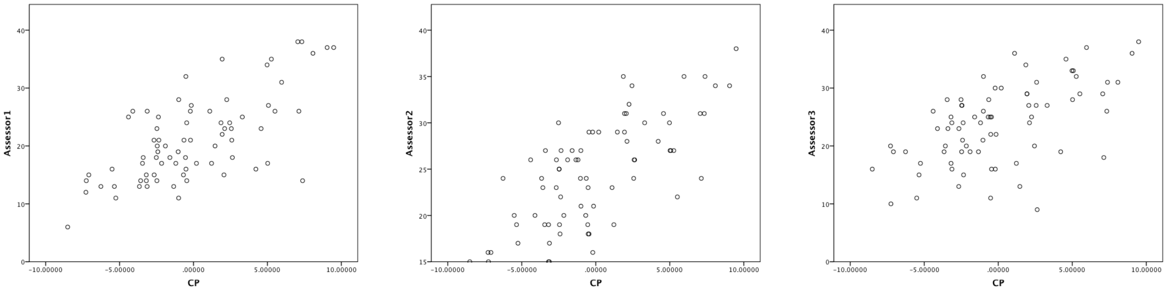


Figure 5.9 Scatter plots between scores from each assessor and Comparative Pairs.

Discrepancy analysis

Statistics from the analysis of the results of Comparative Pairs judgements process this far has shown a high reliability in the scores, a good correlation with the other scoring methods, and a good fit to a Rasch dichotomous model. However, these analyses identified that there were several outlier artworks. These were the portfolios that were scored quite differently to the rest when comparing scores from the different sources. Two differences are discussed in this section; the first is based on the difference of the ranks obtained from the Comparative Pairs judgements method and the Analytical marking, the second is based on the misfit analysis obtained from the ACJ system on the Comparative Pairs judgements method. Ranking and scoring data from each assessor in the Analytical marking were presented in addition to the combined Analytical rank and score to better illustrate the similarities and differences in the portfolios that were different to the others. In both analyses the same method is used; patterns that might emerge from the rankings and scorings were discussed, followed by a discussion on assessors' notes from the ACJ system.

Differences between rankings from Comparative Pairs judgements and Analytical marking

Discrepancy analysis between results of the Comparative Pairs judgements and the Analytical marking was conducted based on the ranks obtained from the two scoring methods. Scores obtained from these methods were in different measurement scales. While the Analytical marking resulted in percentage of raw scores, the scores resulting from the Comparative Pairs judgements were in logits. Consequently, the difference between the two scores given to every student was not meaningful nor comparable therefore the ranks obtained from the scores were used instead. Furthermore, as was discussed in Chapter 2, in the scoring process there are usually variations in the way assessors distribute the scores. For example, the score of 70% given by Assessor A might not represent the same quality as 70% assigned by Assessor B, even if they used the same criteria. This is more pronounced in such subjective tasks. This section looks more closely into those results in order to establish the cause of the discrepancy between the ranks obtained from the Comparative Pairs judgements and the Analytical marking. Ranks from the WACE scores were also discussed as a comparison.

Figure 5.10 depicts the distribution of the absolute differences between the ranks generated by the scores obtained from all three scoring processes, while Table 5.15 shows the descriptive statistics of the absolute differences. The absolute differences between the ranks from the Analytical marking and the Comparative Pairs judgements were quite widely spread with differences ranging from 0.0 to 47.5 for 75 students, with a mean of 11.2 and a standard deviation of 11.3. The absolute differences between the ranks from the Analytical marking and the WACE marking were the least spread with differences ranging from 0.0 to 37.0 with a mean of 9.6 and a standard deviation of 9.0. The absolute differences between the ranks from the WACE marking and the Comparative Pairs judgements had the widest spread with differences ranging from 0.0 to 44.5 with a mean of 13.4 and a standard deviation of 11.5. This further signifies the similarity between scores obtained from the Analytical marking and the WACE marking.

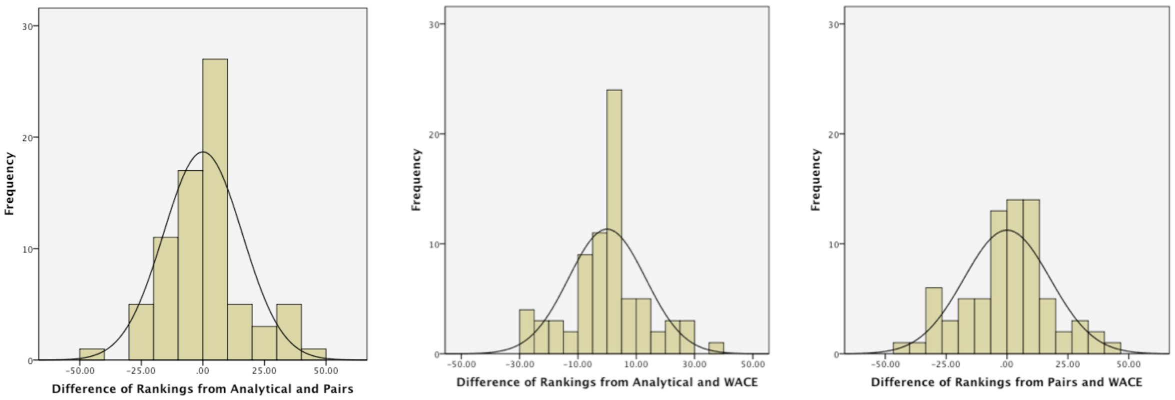


Figure 5.10 Distribution of differences between the rank generated by scores from WACE marking, Analytical marking and Comparative Pairs judgements.

Table 5.15

Descriptive Statistics for Absolute Differences in Ranking Generated by Scores from WACE Marking, Analytical Marking and Comparative Pairs Judgements

	N	Minimum	Maximum	Mean	Std. Deviation
Analytical - Pairs	75	0.0	47.0	11.2	11.3
Assessors1, 2, 3 (largest difference)	75	2.0	63.5	24.6	14.4
Analytical - WACE	75	0.0	37.0	9.6	9.0
Pairs - WACE	75	0.0	44.5	13.4	11.5

A correlation analysis of the absolute difference in rankings from the three scoring processes, including the largest difference between the three assessors in the Analytical marking process, was done to further examine these considerable differences. The correlations are presented in Table 5.16. This analysis indicated that there was not much similarity in the differences between rankings from the scoring processes, except between differences in ranks from the Comparative Pairs judgements and the WACE marking, and between the Comparative Pairs judgements and the Analytical marking with a correlation coefficient of 0.53 (p<0.01). This suggests that there was a moderate degree of similarity between differences in the rankings from Comparative Pairs judgements and the Analytical marking with differences in the rankings from Comparative Pairs judgements and the WACE marking. Consequently, this indicated that the different scoring process and type of criteria could be one of the factors that created difference in the scoring results.

Table 5.16

Correlations Between Absolute Differences in Ranking Generated by Scores from WACE Marking, Analytical Marking and Comparative Pairs Judgements

	Analytical - Pairs	Assessors1, 2, 3	Analytical - WACE	Pairs - WACE
Analytical - Pairs	1	0.13	0.17	0.53**
Assessors1, 2, 3 (largest difference)		1	0.20	0.14
Analytical - WACE			1	0.22
Pairs - WACE				1

**. Correlation is significant at the .01 level (2-tailed).

The lack of strong correlations between the absolute differences of ranks obtained from the three scoring processes further implied that there was no specific consistent procedural reason for the large differences. It indicated that the absolute differences were not caused by differences between scoring methods, which were the difference in criteria, scoring media (i.e., original artwork or digital representations of the artworks), and calculations to obtain the final scores. Consequently, it indicated that the differences were most likely to be caused by factors such as the qualities of the artwork (e.g., creativity, visual language, and materials) the quality of the scoring criteria (e.g., the range of scores, semantics) or the assessors' preference. It should also be noted that statistically the small sample size could also cause the differences between the distance between scores and the distance between ranks, amplifying the distance between scores during the ranking process. This effect is illustrated in a later paragraph.

Because there was no indication that the differences in the rankings were caused by procedural factors in the three scoring processes, the next step was to examine other factors that could cause the difference in the rankings. Artworks with more than 2 standard deviations difference from the mean of the absolute difference between ranks obtained from the two scoring methods were analysed to investigate the possible main reasons for the difference such as the quality of the artworks, the assessor's personal preference, or technical problems. In the Visual Arts course there were five out of 75 artworks (6.7%) with such large difference. Table 5.17 shows the ranks and scores for the five artworks.

Table 5.17

Artworks with More than 2 SD Difference in Ranking in Visual Arts Course

	Rank						Score (%)					
ID	Analytical				CP	WACE	Analytical				CP	WACE
	A1	A2	A3	Ave			A1	A2	A3	Ave		
VQ901	42	31	75	64.0	17	43	45.0	65.0	22.5	44.2	53.4	48.0
VL915	47	58	61	64.0	28	27	42.5	47.5	42.5	44.2	49	54.0
VO906	61	64	64	69.5	34	67	35.0	45.0	40.0	40.0	45.2	35.0
VC902	21	31	29	23.5	67	22.5	62.5	65.0	65.0	64.2	34.8	56.0
VJ904	53	22	53	47.5	15	47.5	40.0	70.0	47.5	52.5	57.8	47.0

With regards to raw scores, there was agreement between the three scoring methods for these artworks except artwork VC902 with a Comparative Pairs judgements score of 34.8%, which was much lower than both the Analytical marking and WACE marking which scored the artworks 64.2% and 56.0 consecutively. For artwork VQ901 there was agreement across scoring methods and Analytical marking assessors except with Assessor 3 which gave the artwork a score of 22.5%, which was much lower than the other scores given to this artwork. Aside from these differences, in general the scores for these five artworks showed only slight differences. *Figure 5.11* shows the scores for these five artworks.

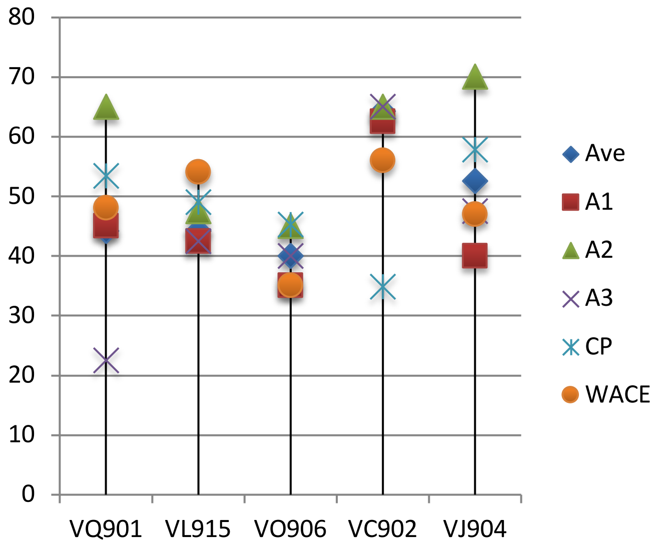


Figure 5.11 Scores for artworks VQ901, VL915, VO906, VC902, and VJ904.

When the score distribution in each scoring method was considered and the scores were converted into ranks, these slight differences between scores from the different methods of scoring, and different assessors in the Analytical marking, were magnified, as is shown in *Figure 5.12*.

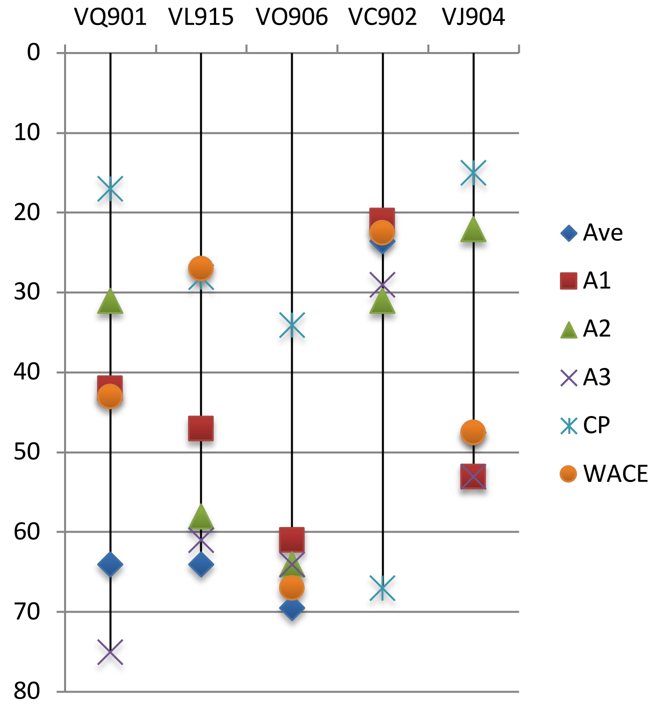


Figure 5.12 Ranks for artworks VQ901, VL915, VO906, VC902, and VJ904.

Artwork VQ901

For VQ901 the scores from the Comparative Pairs judgement and the WACE practical marking were not too different. However, when the score distribution for each scoring was considered in a rank order, the ranks were all different. The rank from the Comparative Pairs judgement was the highest with this work was ranked below the first quartile as the 17th, followed by WACE marking at the 43rd, between the second and third quartiles, and the average of the Analytical marking at the 64th, or above the upper quartile. All three Analytical marking assessors also placed this artwork in three different positions, which were the 31st, 42nd, and 75th.

Notes from the ACJ system in the Comparative Pairs judgement assessors suggested that this artwork was quite simple and incomplete but showed creativity and good technique and skill. The assessors' comments included *concept quite sophisticated in spite of the simple form*, *pity the accompanying sculpture was not included*, *borrowed imagery but somehow the sense of balance and the student's ability to successfully use some of the elements make this a satisfying piece*. An assessor mentioned that he chose this artwork as a winner because of the artist statement was well articulated. Another assessor reported a problem with the quality of the video. There was also a concern over the limitation of the photographs, mainly because they could not capture some of the quality of work in this particular type of medium.

In case of this artwork, the large difference of ranks between the Comparative Pairs judgement and the Analytical marking may have been caused by the contradiction between the simplicity and the sophistication of the artwork in both the technique and the finished product. Scoring methods that used digital representations ranked this artwork very differently, in the first and last quartile. Two of the Analytical marking assessors ranked this artwork quite similarly to the WACE result, however Assessor 3 ranked very differently, even though both scoring methods used the same assessment rubric. It was worth noted that among the three Analytical marking assessors, Assessor 1 tended to utilise the widest range of scores, between 6.0 and 38.0, Assessor 2 the most narrow, between 15.0 and 38.0, and Assessor 3 in the middle, between 9.0 and 38.0.

Artwork VL915

VL915 was ranked closely on the Comparative Pairs judgement and the WACE marking, with ranks that were within the second quartile. These two scoring methods were different in both the types of criteria and marked work. The three Analytical marking assessors ranked the work relatively closely within the second and third quartile, from the 47th, 58th and 61st, for Assessors 1, 2 and 3 consecutively, averaging in the 64th.

Almost all comments on VL915 from the Comparative Pairs judgement indicated that the wins for this work were because the comparison was of a lesser quality. Such comments were *well composed and expresses ideas more clearly than B*, *A* (VL915) *has stronger visual composition. I am making the judgement without being able to see the close ups on B but the composition is less appealing*. This work was submitted without an artist statement and a few assessors regretted that they could not gauge on the depth and meaning of this artwork. Even so, the Comparative Pairs judgement placed VL915 quite high in the ranking.

Artwork VO906

The ranks for VO906 from the three assessors, and hence the assessor average, and the WACE practical marking were all close together in the last quartile. The ranks were 61, 64, and 64 from the three assessors, and 69.5 from the WACE marking. In contrast, the Comparative Pairs judgement placed this artwork in the second quartile, with a rank of 34.

Assessors' comments from the ACJ system indicated that the quality of this work was lacking, especially in the visual composition, painting skill, creativity, and visual communication. Some of their comments were: *Palette is interesting and indicative of the indigenous culture. Lacks skill and appears to attempt too many styles within the one painting. Need to master manipulation and control of brush strokes*, *an idea that beats you around the head, not well composed* and: *corpus conflict*.

The agreement among the three Analytical assessors and the Comparative Pairs assessors' comments did not fit the Comparative Pairs rank of 34, and upon checking, the statistical data from the ACJ system showed that this artwork won four times and lost nine times. As such, there was a concern that the rank was too high for VO906 even though it should be noted that the parameter calculation for the Comparative Pairs judgements was not a simple linear function but instead iterations of numerous probability functions and therefore it was still possible that VO906 had a high rank.

Artwork VC 902

For VC902, the Comparative Pairs judgement ranked the work at the last quartile on the 64th, which was very different to the ranks from the other scoring methods. All three

Analytical marking assessors and result from the WACE marking ranked this work within the second and low third quartile, with ranks ranging from the 21st to the 39th.

From the assessors' comments in the ACJ system, it appeared that this work was considered good but did not compare very well to the comparison works. These comments were ranging from: *pretty ordinary work by sculptural standards*, *some areas are not well manipulated but a good effort overall*, *A resolved work that displays some creativity*, *Although the idea is simple, the resolved work demonstrates a competent application of unusual materials*. Comments on the comparison included: (The other work had) *greater control over media, more sophisticated use of visual language*, (The other work is) *the most 'Art'*, and: (This work is) *…trying to convey something about human relationship with the environment, prompts the viewer to ponder the relationship more than is the case in work B* (the other work). On the Analytical marking Filemaker Pro database one of the assessors included a comment: *time consuming application of leaves and bark – neat and relatively sophisticated*.

For this artwork, the type of work did not seem to make a difference, as the Analytical and WACE markings yielded very similar rankings. The type of assessment criteria, however, might be one of the factors that contributed to the discrepancy between the results from the Comparative Pairs judgement and the other scoring methods. In this case, the holistic criterion might have disadvantaged the student.

Artwork VJ904

For VJ904, the ranks from the Comparative Pairs judgement was similar to the rank given by Assessor 1 in the Analytical marking, each was the 15th and 22nd. The other two assessors placed this work at the same rank of 53, creating an assessor average of 47.5 for all three assessors. This assessor average rank was the same rank as the rank resulted from the WACE practical marking.

The assessors' comments from the ACJ system suggested that this work indicated creativity, good visual communication skills, and good thinking but was lacking in technical skill. In addition to that, for this artwork an assessor predicted a difference between results from the two methods of marking in a comparison with another artwork:

> I suspect that B would have been favored in the [Analytical] marking as it is perhaps more accessible an image and quite eloquent in the composition. The marking key would probably favor the skills demonstrated but for the execution of a creative and original idea A is a more sophisticated artwork.

This was in line with another assessor's comment: *This artwork is a good illustration of the dichotomy between ideas versus skill in the marking key*. *The 'roughness' is however part of the appeal*.

Comparative Pairs misfits

The ACJ system provided data on misfits for both the assessors and artworks. Based on the weighted mean square values (wms) on the artworks, there were four artworks (5.3%) that were above two standard deviations different to the average value. This difference indicated that for these artworks the Comparative Pairs judgement was not consistent, and Pollitt (2012a, 2012b) suggested that because the assessors could not agree on the ranking of such artworks, the artworks should be examined more closely. The disparity could be from the assessors' side or the artworks' side, or both. Several factors could be the reason of this gap, for example assessors' personal preference, a lack of the students' understanding on WACE criteria that were used to develop the holistic criterion, the students' inability to communicate their design, or missing or unaddressed rubric components. Table 5.18 displays these four students' results from the different scoring processes.

Table 5.18

Ranks and Scores for Artworks with Weighted Mean Square (wms) More than 2 SD Difference

ID	Rank						Score (%)					
	Analytical				CP	WACE	Analytical				CP	WACE
	A1	A2	A3	Ave			A1	A2	A3	Ave		
VL904	73	67	61	73	68	70	27.5	42.5	42.5	37.5	40	37.5
VL910	53	64	45	60	36	68.5	40	45	55	46.75	57.5	42.5
VL914	53	54	72	71.5	70	62	40	50	27.5	39.25	40	51.25
VQ904	13	24	7	12	11	20.5	67.5	67.5	82.5	72.5	75	72.5

The raw scores from the three scoring methods for these four artworks were relatively similar. For VL904 only Assessor 1 from the Analytical marking gave this artwork a score that was lower to the others, with a score of 27.5% compared to the other scores between 37.5% and 42.5%. For VL914 only Assessor 3 gave a relatively different score of 27.5%, lower than the scores from other assessors and methods which ranged between 40% and 51.25%. *Figure 5.13* shows these scores.

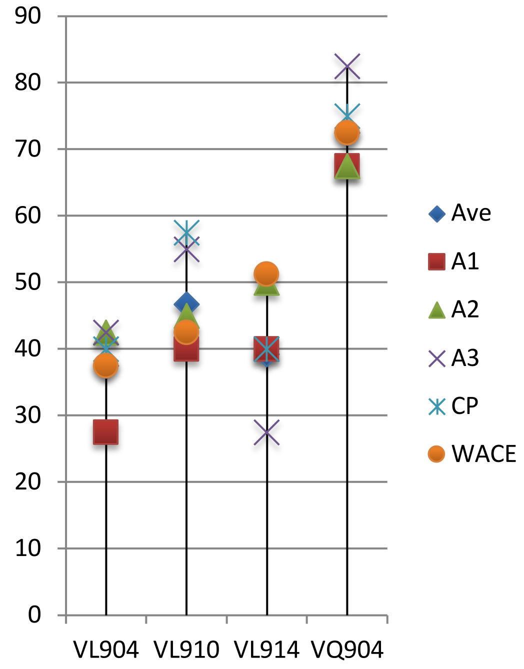


Figure 5.13 Scores for artworks VL904, VL910, VL914, and VQ904.

The rankings for these artworks were not too different across methods and among assessors in the Analytical marking except for artwork VL914. The score from the Comparative Pairs judgements for this artwork was exactly the same with the score from

Assessor 1 with a score of 40%. However, the Comparative Pairs judgements ranked this artwork at the 70th which was similar to the rank from Assessor 3 at the 72nd, while Assessor 1 ranked this artwork higher at the 53rd. Aside from this, there was no apparent differences among scores and rankings given to these artworks. Three of these four artworks were from one school, VL, however further analysis did not reveal any patterns among the three artworks that could cause the inconsistency in judgement. *Figure 5.14* shows these ranks.

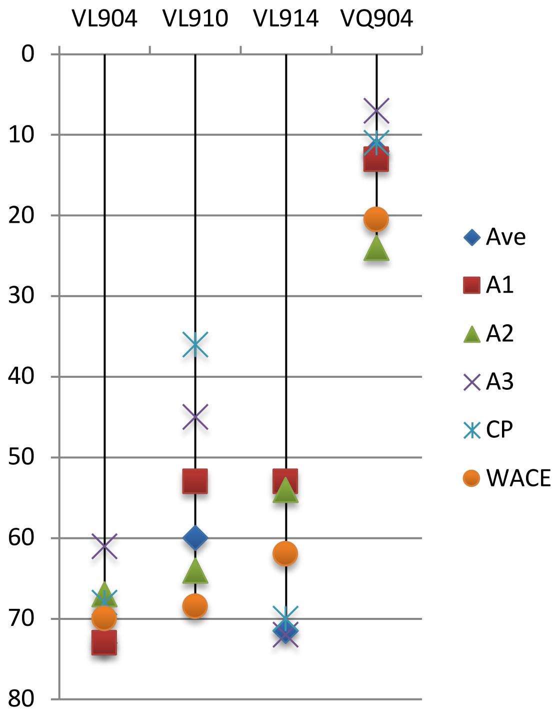


Figure 5.14 Ranks for artworks VL904, VL910, VL914, and VQ904.

Three out of these four artworks were from school VL which was ranked at the lowest in Comparative Pairs judgements and second lowest in the other methods.

Artwork VL904

Artwork VL904 was ranked quite similarly across methods and among Analytical marking assessors with ranks between 61st and 73rd. Even though there was agreement across scoring process, this artwork had a large wms difference to the mean wms from the Comparative Pairs judgements. This indicated that the holistic criterion resulted in inconsistency when it was used to judge this artwork. The Comparative Pairs judgements assessors' comments on the ACJ system were in agreement with the final ranking of this artwork. Such comments were: *both works were on the lower end of the scale*; *superficial and poorly executed*; *lack of creativity and originality, use of materials weak – looks unresolved*; and: *basic skills shown in poster*. There was indication that when compared to a weaker artwork this artwork had some advantage from comments such as: *more evidence of student workings although book cannot be judged* and: *pertinent issues are dealt with although superficially – the inclusion of a sealed book warrants some interest*. The only positive comment that was different to the others and could be a reason for this artwork's wms was: *far more intellectual and quite refined use of materials, much more innovative*.

Artwork VL910

The range of ranks assigned to artwork VL910 was quite wide with the highest was from the Comparative Pairs judgements at 36th and the lowest at 68.5th from the WACE marking. This artwork's scores were relatively more similar with the lowest score of 40% from Assessor 1 and the highest of 57.5% from the Comparative Pairs judgements. Assessors' comments from the ACJ system were mostly comparative for example: *Work of portfolio A* (the pdf file of artwork VL910) *is far superior in all aspect*s, *B* (the artwork being compared to VL910) *is simply the stronger work – A* (VL910) *is decorative but in a lesser league*, and: *Portfolio B A* (the pdf file of artwork VL910) *is a little more original with some experimentation in media*. These comments and the middle range Comparative Pairs judgements ranking were parallel. The misfit statistics for this artwork could be caused by the lack of artist statement accompanying this artwork which left the assessors guessing the intention and media of this artwork, which was suggested by comments such as: *not sure collaged materials on image 2 and 3 correlates with the print* and: *Think it's a print but works quite well*. There was no comment that was too different for this artwork.

Artwork VL914

Both the scores and ranks for this artwork varied with scores ranging from 27.5% from Assessor 3 and 51.25% from the WACE marking, and ranks ranging from the 53rd from Assessor 1 and 72nd from Assessor 2. Assessors' comments from the ACJ system for this artwork all suggested a poor quality of work, for example: *sloppy skills*, *awful*, *manipulation of media is weak and rather clumsy*, *poor execution*, and: *pretty ordinary sculpture… more like yr 10 skill level*. A possible explanation for the large wms difference for this artwork was the Comparative Pairs judgements process in which the comparisons caused this artwork to be judged more superior than the artworks to which it was being compared. The variation in the scores and ranks given by different Analytical marking assessors and the WACE marking suggested there was another possibility that there were Comparative Pairs judgements assessors who considered this artwork had a better quality than the other assessors without leaving a comment on the quality.

Artwork VQ904

Artwork VQ904 was judged to be among artworks with good qualities with scores ranging between 67.5% from Assessor 1 and 2, and 82.5% from Assessor 3 and ranks ranging between 7th from Assessor 3 and 24th from Assessor 2. This artwork was a series of images. Comments from the ACJ system suggested agreement among assessors on the student's Photoshop skill and creativity, as well as on whether all the pieces contributed to the intended image with comments such as: *series of photoshop images skilfully edited, interesting compositions*, *The ideas generated by this student are creative and interesting to look at. The support pieces are not adding to overall impression so perhaps this students would be best served to reduce the number of artworks in the submission*, and: *Images are excellent. Unsure of the quality of supports used*. The contrast between the student's skill and the inappropriate use of supporting images could be the reason the misfit statistics on this artwork.

Assessor Interview

After both scoring processes were concluded, the assessors were asked to give their opinions on the scoring processes, the online tools used, and the quality of the work submitted by the students. 13 of the 15 assessors sent back their responses through email. Assessor demographic data are shown in Table 5.19. The interview consisted of five demographic questions and twelve questions pertaining to the assessors' experience in the scoring processes. These questions are presented in Appendix F.

Table 5.19
Visual Arts Assessor Demographic Data

Assessor	Age Group	Teaching experience (Years)	Teaching VA (Years)	Teach Stage 3 VA in 2011	WACE marker
A	>40	25	25	No	No
B	>40	29	29	No	No
C	>40	15	15	Yes	Yes
D	>40	25	25	-	-
E	>40	17	17	No	Yes
F	>40	25	21	Yes	Yes
G	>40	25	25	No	No
H	>40	19	19	No	Yes
I	20-30	6	6	Yes	Yes
J	>40	21	21	Yes	No
K	>40	25	25	Yes	Yes
L	>40	1	1	Yes	Yes
M	>40	30	25	Yes	No

Twelve out of 13 assessors had at least six years of teaching experience in Visual Arts. One assessor only had one year of teaching experience but was experienced in WACE marking. Eight out of 13 assessors taught Stage 3 Visual Arts course in 2011 and eight assessors were WACE markers in 2011.

The rest of the questions in the interview were designed to gather the assessors' opinion on the quality of the student work, the marking processes and the suggestions they had regarding the complete process. In this study, the assessor interview was used to provide information on the assessors' experience that was likely to influence the validity of the assessment. Therefore, only responses that pertained to issues surrounding reliability and validity are discussed in the next section.

Reliability of Scores

As was discussed in Chapter 3, the ACJ system was designed so that it could continue on creating judgement rounds that would gradually become finer and finer in pairing student work. As this was being done, the reliability of the judgement also became higher, mostly because of the combination of the gradual increase in the number of judgements that consequently increased the cancelling out of the differences between judges and the gradual improvement in the fine-tuning of the pairings. Once the reliability coefficient reached the intended value, when more judgement did not increase the reliability, the judgement session was stopped. In Visual Arts, the judgement session was stopped when the reliability coefficient reached 0.959. This high reliability level reflected both the inter-rater reliability and internal reliability, as calculated by the ACJ system (Kimbell, 2008).

Because of this characteristic, the Comparative Pairs judging was likely to reach a high reliability coefficient, unless the misfits were too bad. The reliability coefficients of the scoring methods were as shown in Table 5.20 There was no reliability analysis available to the researcher on the WACE result. The WACE scores were obtained from double-blind marking and reconciliation between markers. The Comparative judgement reliability coefficient was obtained from the analysis generated by the ACJ system using a statistic analysis similar to Cronbach's alpha coefficient.

Table 5.20
Internal Reliability for Each Set of Scores

Method of marking		Internal reliability
Analytical marking:	Assessor 1	0.934
	Assessor 2	0.915
	Assessor 3	0.934
	Average	0.944
Comparative Pairs judgement		0.959
WACE Examination		n/a

The high internal reliability specified by the Cronbach's Alpha coefficients obtained from the SPSS software for the Analytical marking represented the internal reliability of the criteria. These reliability coefficients indicated that there was an overall agreement among the criteria in the rubric. The inter-rater reliability was represented by the correlation between assessors. For Visual Arts the correlation coefficients between assessors ranged from 0.51 and 0.56 ($p<0.01$) and between 0.49 and 0.56 ($p<0.01$) respectively for the score and rank, as was shown previously in Table 5.13. These coefficients indicated that there were only moderate correlations between assessors in the Analytical marking. In summary, the internal reliability among criteria was high but the correlation between assessors was only moderate. In contrast, the reliability coefficient of the Comparative Pairs judgement, which represented both the internal reliability, or internal consistency in judgement, and the inter-rater reliability, was high.

While the reliability coefficient in the Analytical marking meant that the criteria in the assessment rubric measured similar constructs, which was only one factor of the reliability measure of the marking result, the reliability coefficient of the Comparative Judgement included both types of reliability.

Validity of Assessment

Three points of reference are used to discuss the validity of the Comparative Pairs judgement. The first is from the reliability of the result of judgement, then from the way the result was compared from results from the other scoring methods, and lastly, from the issues that might threat the validity of the result as were disclosed by the assessors in the interview.

Reliability of result

The result from the Comparative Pairs judgement had a high reliability coefficient, therefore the threat from the lack of both the internal and the inter-rater consistency could be disregarded. As Pollitt (cited in Kimbell et al., 2009, p. 79) posited, in Comparative Pairs judgement, variation in both the absolute standard and the weightings did not influence the validity of Comparative Pairs judgement result. One factor that might have an effect on the validity of this result was the variation of the artwork that was judged to be the winner because of the judges' different perspective. The internal reliability in the Comparative Pairs judgement result was 0.936, which reflected a high confidence in the consistency of the judgements. The assessor misfit statistics also did not show a problem with this.

In contrast, for the Analytical marking, the internal reliability was high but the inter-rater consistency was only moderate. This indicated that even though the criteria measured the same set of skills, there was inconsistency in how the assessors used the rubric.

In both scoring processes, only experienced assessors were selected. This was aimed to avoid differences among assessors that were caused by lack of experience. Technical help was also provided in both processes to avoid disturbance by technical problems such as difficulties in accessing the interface. For the Comparative Pairs judgement the holistic criterion was discussed together by most assessors based on the WACE examination criteria. This was aimed to avoid differences in understanding the holistic criterion. These efforts were taken as a precaution to limit the factors that could potentially compromise the validity of the result.

Comparison with results from other scoring methods

Correlations between scoring methods (Table 5.13) indicated that results from the Comparative Pairs marking was significantly and moderately correlated with results from both the Analytical and the WACE practical markings with correlation coefficients of 0.79 and 0.74 consecutively for the scores and 0.73 and 0.67 for the rankings.

In contrast to that, in the Analytical marking, even though the internal reliability level of each assessor was high, the low correlations between assessors lowered the confidence of the validity of the result of marking. Averaging the results from the three assessors did moderate the results but that still did not quite bring the confidence level in the results to the same level as the results from Comparative Pairs judgement.

Validity issues emerging from the assessor interview

Regarding the quality of the digital representation of student work in Visual Arts, the assessors reported dissatisfaction. None of the 13 assessors considered the digital representation as adequate in demonstrating the original student work. The resolution of the photographs and videos was reported to be too low, with comments such as: *Many of the digital images were blurred. The segments/enlargements did not really help as sometimes they appeared to be from the original photo i.e. had the same pixilation* [sic] and: *some of the images were blurry or did not show sufficient detail*. Assessors also conveyed their disappointment over the video quality with comments such as: *The videos were not of much use to me which was a pity as I thought the sculpture could have been explained better using this method* and: *Video footage was often wobbly and swinging to and fro*, however an assessor felt that the videos still provide: *an indication of size*. Several assessors considered the quality of the digital representations to be: *some better than others* and: *a little hit and miss at times*, conversely, one assessor said that even though she thought the quality was poor, *the only positive is that all photographs were of the same quality*.

Beside the quality of the photographs and videos, there were also concerns over the lack of depth, texture and clarity of the original work in the images, different colour with the original, as indicated by comments like: *I felt quite removed from them*, *there were several pieces of work that I had marked in the WACE and were outstanding – and I felt that for many of these works this was not communicated in the digital format*, *Texture of 2 and 3 dimensional artwork difficult to discern. … subtlety was difficult to discern*, and: *despite having a matchbox as an indicator of size – scale and dimensions still are unclear*. The condition of the digitisation of artwork was also considered to be less than ideal, with assessors commenting: *the works should have been hung not leaning on an easel where the angle of inclination was distracting some of the work*, *inadequate lighting conditions altered the colour palette of the artworks and created flare*, *works made up of multiple pieces did not come though* [sic] *in a unified way*, and: *too much interference from surrounding artwork*.

Understandably, the assessors' low view on the quality of the digital representations led to issues regarding reliability and validity*: I feel that in some instances the students were at a distinct disadvantage as a result of the poor quality of the visual material*. Furthermore, beside the possibility that the poor representation might disadvantage good quality work, it was also considered possible that poor quality photographs could over-represent certain artworks as implied by an assessor: *alternatively, the photos often complimented an artwork reducing faults that were easier to see in real life*.

Regarding the types of artworks that might be represented well in digital representations, most assessors doubted that either two-dimensional or three-dimensional artworks could be well represented digitally, however, there were several suggestions including using professional studios; better equipment such as tripod, lighting, neutral background, and lenses; and more focussed close-ups. Regarding the way the digital representations were displayed, they mostly preferred the PDF files, picture files, and the PowerPoint files.

As was discussed earlier in this chapter, there were two online systems that were used for scoring: the ACJ system used for the Comparative Pairs judgements and the Filemaker Pro system used for the Analytical marking. All Visual Arts assessors found these systems to be working well and easily accessed. A few instances when the systems were lagging or the internet connection was slow were reported, however overall there was no technical issues related to the scoring system. Consequently, the scoring system was not considered to be a threat to the validity of the assessment.

The Comparative Pairs judging process was considered to be easy by most assessors. Difficulties were reported to be caused by artworks that were similar, as one assessor informed: *mostly easy but some works were very similar in quality and those took some time as it was not always easy to discern how more competent the student was in skills and techniques and application of paint etc*. Several assessors found that the Comparative Pairs judgements method with a holistic criterion to be preferable than the Analytical marking because it was easier and more suitable for the Visual Arts course, especially when the quality of the digital representation was low. However, several assessors were worried over the quality of their judgements, such as revealed by an assessor: *it was easy* *to make a judgement, whether that judgement was accurate due to the filter I viewed the work through is another question*. Two sources of inaccuracy were mentioned, which were the judgement method as in: *I am dubious that this is a fair and consistent means of making comparisons between student's work* [sic] and the quality of the digital representations as in: *some were difficult because the skills quality was hard to determine from photo or no sense of scale*.

When asked about the overall quality of the artworks, there was a range of responses. Most assessors considered the quality of the artworks was quite average or ranging from low to high. Two assessors considered it to be below average while another two assessors thought rather highly of the works.

Assessors who were involved in the WACE marking and thus have seen and marked the original artworks considered the experience as affecting their judgements. Other factors that could have influenced their judgements were the quality of the digital representations as reported: *the video file and poor picture quality actually interfered with my appraisal of the work* and: *how well it was presented and whether I got a sense of the overall artwork from the photos*, and: their education as indicated in *I try to take into account contemporary conceptual and aesthetic sensibilities. … some of my contemporary art educators are worryingly dismissive of anything post 1850*.

Summary

This chapter presented the data analysis from the Visual Arts course. Chapter 6 presents the comparison between findings from the Design and Visual Arts courses followed with the Discussion section.

CHAPTER 6
CROSS-CASE ANALYSIS AND DISCUSSION

This chapter reports on the analysis of the similarities and differences in findings between the two courses, Design and Visual Arts. These courses have been treated as two cases of assessment of creative work with the nature of the work and the digital representations being very different. This analysis is structured using the conceptual framework. While Chapters 4 and 5 discussed findings in each course separately, by focusing on the comparisons between scoring methods, this chapter considers how the differences in the nature of the assessment tasks in the two courses might influence the validity and reliability of the results from the Comparative Pairs judgements. Findings from this chapter provide information on how different types of tasks would benefit from the use of Comparative Pairs judgements as well as the limitations of the judgements in the different types of task. The structure of this chapter is similar to Chapters 4 and 5 followed by a discussion directly related to the research questions for the present study. The chapter begins with the assessment task and then the task assessment.

Assessment Task

Both courses investigated in this study had a major practical component, however, the type of this practical component was different. This section discusses the similarities and differences of the nature of these practical tasks in the two courses in relation to the factors that could influence the quality of the scoring results, especially the scoring results from the Comparative Pairs judgements. These factors are the nature of student work, the constraints from the digitisation processes, and the technical limitations. This section is concluded with a discussion on the way these factors could affect the validity of the assessment.

Nature of student work

In Design, the student work was a 15-page portfolio. It contained components that displayed evidence of the Design process of up to three Design projects. These components could be pictures, descriptions, sketches, schemes, mind maps, and photographs that showed the evidence of each student's projects. In Visual Arts, the student work was a finished artwork that could be two dimensional (2D), three dimensional (3D), or motion and time-based. It also included an artist's statement and photographs to indicate how the art should be presented.

The process in the Comparative Pairs judgements for Design was comparing the overall quality between two digitised portfolios presented as a PDF, based on a holistic criterion. The assessors had to examine the details in the 15-page portfolios to decide the winning portfolio in each pairing. Consequently, the challenges in the Comparative Pairs judgements in the Design course mostly stemmed from the numerous components combined with the length of the portfolio. Furthermore, because this scoring process was holistic, there was a possibility that the visual presentation of the portfolio skewed the assessors' judgements, especially when the two portfolios being compared were of similar quality.

The Comparative Pairs judgements process in Visual Arts course was quite straightforward because it assessed a final product based on a holistic criterion. The main problem in this process was from the variety of the types of submitted artworks. In Visual Arts, the WACE examination practical component could be in the form of paintings, sculptures, printed works, and many others. Comparing two artworks that were very different in nature could be challenging and highly subjective. It depended on the assessor having a good understanding of the standards for judgement and experience in making these judgements. However, unlike for Design, they did not have much reading to do; the information was largely visual.

The comparison process in Design and Visual Arts had different challenges that were related to the nature of student work. In Design, the submitted portfolios had differences such as in the contexts of the portfolio (e.g., photography and technical graphics), portfolio materials, and finish (e.g., combinations of sketches and written explanations). Nevertheless, the digitised portfolios were all similar: 15 pages of digital portfolios. In Visual Arts, conversely, there were variations in both the submitted works and the digital portfolios presented in the online scoring processes. The submitted work could be two-dimensional or three dimensional; the artworks could comprise a single piece or several components; they could be drawings, prints, sculptures, or others; and they could be in different sizes below 2.5 m^2 for two-dimensional artworks, and 1.5 m^3 for three-dimensional artworks. The digital portfolios presented for the assessors in the online scoring process contained an artist statement, installation photograph, one full image for two-dimensional works and four to five for three-dimensional, four close-ups, one video for two-dimensional works and large three-dimensional, and an additional video for small three-dimensional works that could be fit onto a revolving table.

In Design the assessors had to compare the overall quality of details spread across 15 pages of portfolio and they would have to mentally sample from written explanations and images, while in Visual Arts the assessors had to compare finished products. These challenges were likely to affect the reliability of the scoring result, and consequently its validity, in different ways. These issues will be discussed later in this chapter.

Constraints from the digitisation process

In general, constraints from the digitisation process in both courses were quite similar with time limitation being the most problematic constraint. The digitisation process had to be completed within two days for Design and one day for Visual Arts. In some cases this time limitation did not allow for adequate problem solving and quality control. In Design there were problems with scanning portfolios that were submitted in materials that were too thin, too thick, or glossy. These portfolios had to be scanned manually and even so could still result in digital portfolios with a difference in colour, brightness, or clarity to the original paper portfolios. Because of the time constraints, in some cases compromises had to be made.

In Visual Arts, there were artworks that were difficult to install because of their dimensions and difficult to capture digitally because of their dimensions or materials. The use of equipment that could improve the quality of the digital representation of student work such as lightings and backdrops was made impossible by the time constraint. There was also very limited time available for ensuring the quality of the photographs and videos sufficiently represented the original artwork. Problems encountered during the digitisation process might reduce the quality of the PDF portfolios in Design and the photographs and videos in Visual Arts, which in turn could affect the results of the scoring processes. There was evidence of this that is discussed later. In general constraints from the digitisation process were probably less of a concern for Design than for Visual Arts because the number of portfolios that were difficult to scan were not that many and it was the only source of problems. In Visual Arts, the problems arisen from different sources such as lighting, installation, and the dimension of the artworks.

Technical limitations

In Design there were technical limitations due to the difficulty in scanning certain types of material such as glossy paper, as well as from the file size of the PDF files of the scanned portfolios. Because the portfolios contained 15 pages of A3-sized paper, the size of the PDF files was quite large, around 15 MB. In Visual Arts, the variety of the dimension, type, and material of the artworks, as well as the size of the photographs and videos caused the technical problems. The digital representations of the student work should represent the original artwork as closely as possible; hence the size of the digital files was quite large. Large file sizes could be a problem during online scoring process especially when the assessors' Internet connection was slow, consequently it could affect the results of the scoring processes. There was evidence of this that is discussed in the next section. Similar to the constraints from the digitisation process, in general, the technical limitations in Design also was not as influential as in Visual Arts because in Design there was only one PDF file.

Discussion based on Assessment Task

The nature of the student work, the constraints from the digitisation process, and the technical limitations were the factors related to the Assessment Task identified to affect the overall quality of the assessment. In this study, students' original works were digitised and uploaded to the servers for the online scoring processes. This Assessment Task part of the assessment processes influences the judgement processes which in turn influences the judgement results. This discussion is based on findings from the assessor interviews.

In the Comparative Pairs judgements in the Design course, the student task was in the form of a 15-page PDF file containing written and image works. While there were limitations that could reduce the fidelity of the scanned file compared to the original paper portfolio as well as technical limitations that could slow the judgement process, the assessors did not consider the quality of the digital representations and the process to be problematic. The digital files were reported to be mostly clear and easy to access, therefore the quality of the digital representations of student work in Design was not considered to affect the reliability of the scores and the validity of the assessment.

On the other hand, in Visual Arts, the quality of the digital representation of the students' artworks was reported to be low. The Visual Arts assessors reported that the photographs and videos of students' artworks did not represent the artworks well, with details such as layers, textures, colours, and media indiscernible. Consequently, there was a lack of confidence among assessors over the judgement results due to this concern.

Regarding the judgement process, Design assessors were divided on their attitude towards the Comparative Pairs judgements method. The use of a rubric in the Analytical marking process tended to make this process easier in the Design course because the rubric guided the assessors on what to look for in the portfolio as well as on the score range for each mastery level in each criterion. Assessors who preferred the Analytical method considered this method to be easier and would provide accurate results. However, assessors who preferred the Comparative Pairs judgements also considered the holistic judgement to be easier, accurate, and more objective. These assessors found it easy to memorise the holistic criterion and judge the winner of each pair and regarded that variations in judgements would be cancelled out by the number of assessors and judgements.

In Visual Arts, the Comparative Pairs judgements method was preferred by most assessors because it was found to be easier, objective, and more suitable to the nature of the course. Even so, most assessors still reported that they had doubts over the accuracy of their judgements. Their doubts stemmed from their uncertainty over the fairness of this scoring process and their dissatisfaction of the quality of the digital representations.

Issues related to the assessment task such as the quality of the digital representations were viewed by assessors as a possible threat to the reliability of the scoring results and the validity of the assessment. In Design, the quality of the digital representations were reported to be sufficient, unlike in Visual Arts, with assessors reported their dissatisfaction with the quality of the photographs and videos. Aside from this issue, while several assessors from both courses considered the Comparative Pairs judgements method to be an objective and reliable method, others were concerned about the fairness and validity of the Comparative Pairs judgements. Results from the task assessment in both courses, however, indicated good reliability. These results are discussed next.

Task Assessment

This section presents the comparisons of Design and Visual Arts based on data pertaining to the scoring results from the Analytical marking, the Comparative Pairs judgements, and the official WACE marking. Scoring data, including the time taken for scoring, was recorded in the online scoring interfaces in Analytical marking and Comparative Pairs judgements processes. In Analytical marking there were two Design assessors and three Visual Arts Assessors, while in Comparative Pairs judgements process there were 10 Design assessors and 15 Visual Arts assessors.

Scoring time

The average time taken for judging each portfolio online in Comparative Pairs judgements did not vary much between Design and Visual Arts, as is shown in Table 6.1. The types of the original student work in Design and Visual Arts course were different, as well as the digitised version of student work in the two courses; therefore the judging procedure in the two courses were also quite different. In the Comparative Pairs judgements procedure in the Design course, the assessors were presented with a pair of PDF files viewed side by side. Using the holistic criterion, the assessors examined the 15 pages of the PDF files and made a judgement on which portfolio was more superior. In Visual Arts, the assessors were presented with a folder containing a PDF file and a PowerPoint file containing the artist statement, an installation photograph, four close-ups, one full photograph for two-dimensional work or five full photographs for three-dimensional works; the individual image files; and videos; from which they could choose to view. Because of the ACJ system limitations, the content of the Visual Arts files could not be presented side by side; therefore the comparing process between the courses was different.

In Design, the time needed by the Comparative Pairs assessors to judge a pair of portfolios ranged from 2.53 to 11.21 minutes, averaging 4.64 minutes per portfolio, while in Visual Arts the range was between 2.22 and 9.18 minutes, averaging also 4.64 minutes per artwork, coincidentally the same value to two decimal points. While it was expected that in Visual Arts the judgements might take more time, this was not the case. This might be due to a combination between the *chaining* that occurred after the sixth round in the Comparative Pairs judgements and the type of task in Visual Arts. In Visual Arts, the student work was a finished product, which qualities were easier for the assessors to remember. Because after the sixth round, the pairs presented to the assessors contained one artwork that had been compared in the previous pair, the assessors only needed to examine the second artwork, which saved time. Conversely, the Design portfolio contained many elements; therefore the qualities were more difficult to remember.

Table 6.1

Scoring Time for Design and Visual Arts

Scoring Process	Design	Visual Arts
Comparative Pairs judgements	2.53 to 11.21 minutes per judgement Total: 507 judgements in over 39.2 hours Average: 4.64 minutes per judgement* 10 assessors	2.22 to 9.18 minutes per judgement Total: 435 judgements in 33.6 hours Average: 4.64 minutes per judgement* 15 assessors
Analytical marking	Total: 82 portfolios in 17.5 hours 6.4 minutes per portfolio, ranging from about 5 to 15 minutes for each portfolio 2 assessors	Total: 75 artworks in 37 hours 9.9 minutes per artwork, ranging from about 5 to 20 minutes for each artwork 3 assessors

Note: * incomplete judgements excluded.

The Analytical marking assessors took between 5 to 15 minutes to score a Design portfolio, averaging on 6.4 minutes per portfolio, and between 5 to 20 minutes per Visual Arts work, averaging on 9.9 minutes per artwork. In the Comparative Pairs judgements the judging time for both courses was similar while in the Analytical marking the scoring time for Design was lower than Visual Arts. This might be due to the reported issue with internet connection speed in downloading the photographs and videos in Visual Arts and the Analytical marking process that required more details than the Comparative Pairs judgements.

Figure 6.1 graphically depicts the scoring time in the Comparative Pairs judgements and Analytical marking for each course. In each course there was a substantial difference of judging time between the Comparative Pairs judgement and the Analytical marking processes, which could be caused by differences in the two processes and the comprehensiveness of the criteria used in the processes. This difference was more pronounced in Visual Arts than in Design. One of the possible reasons was the time it took to download the photographs and videos in Visual Arts. However, in Comparative Pairs judgement in Visual Arts the assessors needed slightly less time than the Design assessors, even with the videos. Since the different scoring interfaces used in the two processes worked quite similarly, it was likely that the interfaces were the reason for this; therefore the more plausible reason for the difference was the use of the rubric in the Analytical marking. The Analytical marking rubric required the assessors to make a series of judgements rather than just one.

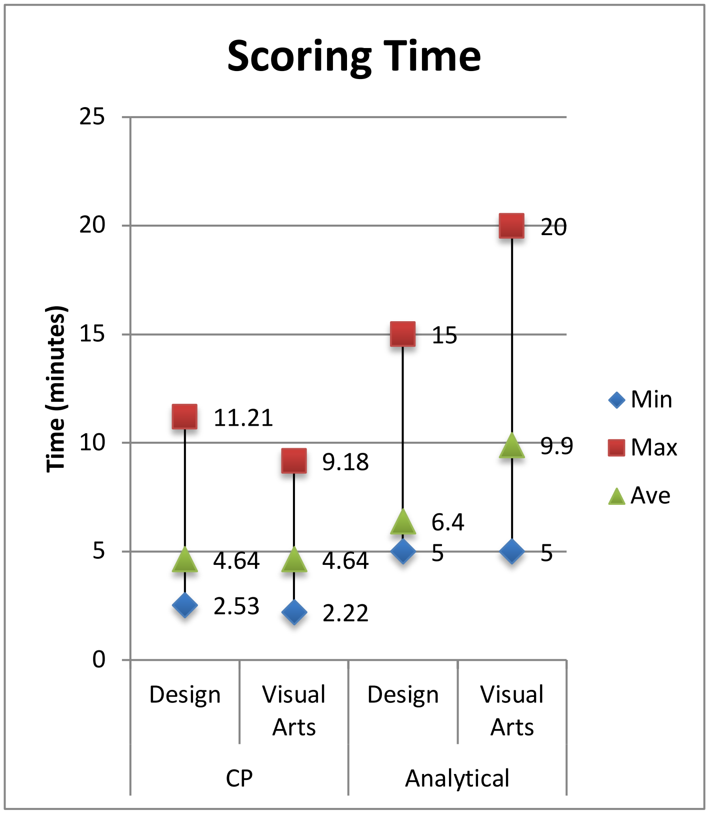


Figure 6.1 Scoring time graph.

In the Design course the file size was relatively similar for each portfolio, hence the loading time for each portfolio should be relatively similar as well. Consequently, longer scoring time could indicate portfolios that were difficult to score. In Analytical marking, the reasons could be associated with difficulty in matching the qualities of the portfolios with the corresponding criteria and descriptors, a lack of clarity in the PDF files, or contradicting qualities within portfolio. In Comparative Pairs judgements, longer scoring time could indicate difficulty in deciding the better portfolio based on the holistic criterion, which could happen when the two portfolios had similar qualities or the qualities were contradictory, or when the assessors had problems in finding the components upon which they could base their judgements, similar to in Analytical marking. In Visual Arts, the file size varied, hence variations in scoring time was inconclusive as longer scoring time could indicate difficulty in scoring, longer loading time, or both.

Analytical marking

The Analytical marking was conducted online by using a Filemaker Pro scoring interface developed for this purpose. The analytical marking rubric used in this marking was the official rubric that was used in the WACE practical examination. In Design, there were six criteria with maximum score points ranged from 6 to 10 for each criterion with a total score of 50. In Visual Arts there were five criteria with maximum score points ranged from 5 to 12 for each criterion, with a total score of 40. Scores from the Analytical marking were analysed based on schools and on assessors to examine possible patterns that might characterise individual schools and individual assessors.

Analysis of scores based on schools

When the mean score for each criterion for each school was calculated, there was no school that had a mean score that was substantially different from each criterion mean score. All schools had mean scores that were no more than 2 SD's difference to the total mean in each criterion in both Design (Table 4.3) and Visual Arts (Table 5.3). In Design these mean scores ranged between 58.0% and 65.5% while in Visual Arts they ranged between 54.6% and 63.3%. This indicated that the criteria contributed reasonably similarly in each school in both courses.

For both Design and Visual Arts there was a distinct pattern on the mean score for each school in each criterion, which could be seen in *Figure 4.1* and *Figure 5.1*. These graphs depict each school's mean score in each criterion, showing that the relative positions of each school compared to one another across criteria tended to be consistent. Schools that scored well in one criterion tended to score well in the other criteria. This pattern was stronger in Visual Arts than in Design. The pattern could indicate good consistency between criteria as well as the influence of school culture, such as teaching style and collective academic characteristics on student performance.

Analysis of scores based on assessors

The analysis of scores based on assessors in both courses showed that the correlation between assessors were significant, but relatively low, with a correlation coefficient of

0.53 (p<0.01) between the two Design assessors and between 0.51 and 0.56 (p<0.01) between the three Visual Arts assessors. There were variations in scores among assessors in both courses in terms of score range and spread, however there were only a few student works that were scored extremely different. In Design there were only three out of 82 portfolios (3.7%) which had a difference more than two standard deviations to the difference mean and there were only three out of 75 (4%) such artworks in Visual Arts. The mean of the difference in scores in Design was 5.6 (SD=4.4) and it was 8.5 (SD=4.3) in Visual Arts.

Analytical and WACE practical markings

When compared to the WACE scores, the means of the Analytical marking scores in both courses were numerically lower than the WACE means, as is shown in *Figure 6.2*. In Design course the difference between the means was more noticeable than in Visual Arts, but in both courses it was still less than a half standard deviation. In both courses the Analytical markers did not utilise the maximum scores, unlike in WACE practical marking. In Design the Analytical marking assessors tended to give lower scores than the WACE markers. In Visual Arts the Analytical marking score range was narrower than the WACE score range.

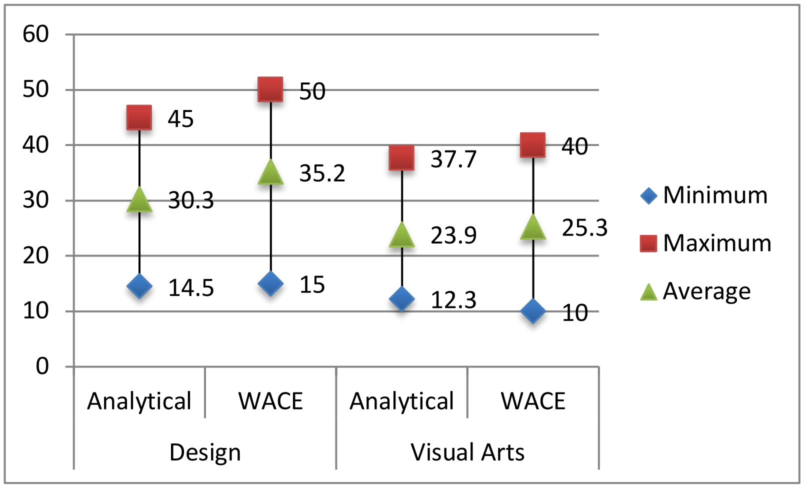


Figure 6.2 Comparison of score range between scoring methods in each course.

Between Design and Visual Arts there was a difference in the number of Analytical marking assessors. There were two assessors in Design and three in Visual Arts. Since more assessors generally could better moderate scores, it was possible that the mean

Analytical scores in Visual Arts correlated better with the WACE scores than in Design because there was one more assessor in Visual Arts.

In both courses correlations between individual Analytical marking assessors were significant and low-to-moderate. Conversely, the correlations between individual marking assessors' scores with the WACE scores were better in Visual Arts than in Design. In Visual Arts, the correlation coefficients were 0.70, 0.71, and 0.75 ($p<0.01$), while in Design the coefficients were 0.36 and 0.55 ($p<0.01$), as was presented in Tables 4.12 and 5.13. This could mean that even though there was a possibility that the different number of assessors in Design and Visual Arts could be the reason the mean Analytical scores and the WACE scores correlated better in Visual Arts than in Design, there could be other factors such as the quality of the marking rubrics.

Considering Design was a new WACE examination course, the Design marking rubric was less tried than the Visual Arts marking rubric. Components of Design marking rubric such as the weighting and the score range for each criterion could affect the reliability of the marking scores.

Comparative Pairs judgements

The Comparative Pairs judgements involved portfolios from 82 Design students from six schools, 10 Design assessors, portfolios from 75 Visual Arts students from ten schools, and 15 Visual Arts assessors. There was not enough data from one Design assessor and one Visual Arts assessor, therefore data from them were excluded.

Data analysis from the ACJ system suggested that in general there was agreement among assessors and good consistency in judgements for both Design and Visual Arts. There were only 25 out of 543 judgements (4.6%) that were considered inconsistent in Design, and 42 out of 497 judgements (8.5%) in Visual Arts. Scoring data also did not indicate extreme misfits in student location parameter, with only six out of 82 Design portfolios (7%) and four out of 75 Visual Arts works (5%) having weighted mean square values that were more than two standard deviations different from the means. The normality test showed that the Design scores were not normally distributed but symmetrical and not significantly skewed while the Visual Arts scores were normally distributed, symmetrical and not significantly skewed. This served only as a description of the score distribution since the Rasch dichotomous model does not require a normal distribution.

Summary of scoring results

The summary of scoring results from the three scoring processes indicated that for most schools, the schools' mean scores from the Comparative Pairs judgements were mostly the lowest than the means from the other two methods in both Design (*Figure 4.5*) and Visual Arts (*Figure 5.5*). In both courses there was a pattern that suggested that there was consistency in the schools' score means from the three scoring processes. This pattern was also apparent in the analysis of data from the Analytical marking (*Figure 4.1* and *Figure 5.1*), in which there was consistency in the schools' score means in each criterion. This indicated that there was relative consistency among the results from the three scoring methods, or that there could be typical academic characteristics in each school that might influence student achievement relative to students in other schools. Consistency among results from scoring is further discussed below using correlation analysis, while the possibility of typical academic characteristics in each school is not discussed further. The pattern from the three scoring processes was more pronounced in Design than in Visual Arts while the pattern from the criteria was more pronounced in Visual Arts. This could indicate that in Visual Arts the rubric generally measured the same construct, relatively more so than in Design.

Comparison between results from scoring methods

The correlation analysis between scores from the Comparative Pairs judgement and both the Analytical marking and the WACE practical examination results for the two courses were both moderate and significant (Tables 4.12 and 5.13). In Design, these correlation coefficients were 0.63 ($p<0.01$) between the Comparative Pairs method and Analytical marking, and 0.67 ($p<0.01$) between the Comparative Pairs method and WACE practical marking. In Visual Arts, they were 0.79 ($p<0.01$) between the Comparative Pairs method and Analytical marking, and 0.74 ($p<0.01$) between the Comparative Pairs method and WACE practical marking. This indicates that the scores from the Comparative Pairs judgement result were quite similar to the results from the other two scoring processes for both courses. This consequently means that in this particular analysis, there was no indication that the following factors created large differences between results:

- the difference between the type of task in the Design and Visual Arts courses.
- the difference between the scoring media in the three scoring processes (i.e. original work or digital representations), both in Design and Visual Arts.
- the difference in the types of criteria that were used to base the scoring on in the two scoring methods for both courses.

The correlations between the average Analytical scores and the WACE scores were moderate for Design (r=0.52, p<0.01). Between Assessor 1 and WACE the correlation was also moderate (r=0.55, p<0.01), it was low between Assessor 2 and WACE (r=0.36, p<0.01), and moderate between assessors (r=0.53, p<0.01). This suggests that the correlation between the average Analytical scores and the WACE scores was moderate only because the average Analytical score was the average of the scores from the two assessors who gave quite different scores. The correlations between individual assessors and the average were understandably similarly high (r=0.89, p<0.01).

In Visual Arts, the correlation between the average Analytical scores and the WACE scores was high (r=0.86, p<0.01), with the correlations between each of the three assessors and the WACE scores all moderate to high (r=0.70, 0.71, 0.75, p<0.01), even though the correlations between assessors were only low to moderate (r=0.51, 0.54, 0.56, p<0.01). This suggested that there were variations in scores among assessors with each correlated quite well with the WACE scores. Perhaps the three assessors tended to look at different things but the average cancelled out the differences. The correlations between individual assessors and the average were understandably similarly high (r=0.80, 0.81, 0.85, p<0.01).

The correlations between the Analytical scores and the Comparative Pairs scores showed a similar trend with the Analytical scores and the WACE scores. In Design Assessor 1’s scores moderately correlated to the Comparative Pairs scores while Assessor 2’s scores had only a low correlation to the Comparative Pairs scores. In Visual Arts, on the other hand, even though there were variations in the correlation coefficients between each assessor and the Comparative Pairs scores, the coefficients were all in the moderate range. The average Analytical scores in both courses moderately correlated to the Comparative Pairs scores.

In summary, correlations between methods of scoring in the two courses suggested that while in Visual Arts there was relative consistency among assessors and scoring methods; that was not the case in Design. This suggested that either in Design, there was a problem with either Assessor 2, the marking rubric, or the difficulty in sampling information in portofolios. In addition, the moderate correlation between both the Comparative Pairs judgement results and the Analytical marking results with the WACE results indicated that there was no evidence of a significant difference between digital and original works to the results of scoring.

Discrepancy analysis

Discrepancy analysis in Design did not indicate procedural factors as the main reason for discrepancies in scores. Even though there were major procedural differences between scoring processes; such as type of criteria, scoring method, and type of scoring media; there was no evidence that suggested that these differences caused the discrepancies. The main reason for the discrepancies seems to be assessor preference. In Design, there seemed to be a propensity among assessors to lean either towards “process” or “product” when they judged a Design portfolio.

In Visual Arts, the discrepancy analysis suggested that there was some degree of similarity between artworks that were scored differently in Comparative Pairs judgements and Analytical marking with artworks that were scored differently in Comparative Pairs judgements and WACE. The low but significant correlation between the differences indicated that the aforementioned procedural factors could cause a difference in the results between the Comparative Pairs judgements and other scoring methods. Notes from the ACJ system suggested that in Visual Arts the assessors balanced their judgements based on different factors within the criteria, for example inventiveness, visual communication, and skills. When a holistic criterion was used, the lack of weighting and score range could push the assessors to become either balanced or biased towards several factors in using their own mental weighting. In Comparative Pairs judgements process, closely similar pair of work could strengthen this tendency.

Differences between Comparative Pairs judgements and Analytical marking

Comparisons between rankings obtained from different methods, and from different assessors, in the Analytical marking showed that both in Design and Visual Arts the absolute differences of ranks between scores from assessors in the Analytical marking were the largest. While some differences in ranking were large to some extent this was due to the relatively small sample. Small changes in scores can lead to big differences in rank. In Design the differences ranged between 0.0 and 76.5 places while in Visual Arts the range was between 2.0 and 63.5 places. These differences highlighted the influence of personal judgement in the ranking of student work even with the use of analytical marking rubrics, which was slightly more prominent in Visual Arts. However, this could possibly be offset by moderation among assessors.

Correlations between differences in rankings obtained from different scoring processes in Design showed there were significant but low correlations between the Analytical-CP differences and the Analytical-WACE differences, and the latter with CP-WACE differences in Design. There was no significant correlation between other differences. This indicated that procedural factors were not likely to cause large differences between results from different processes. In Visual Arts there was a significant but only low-to-moderate correlation between the Analytical-CP differences and the WACE-CP differences, and no significant correlations between other differences. This indicated that in Visual Arts procedural factors were relatively more likely to cause the differences, beside non-procedural factors such as artwork quality and assessor subjectivity.

There were four out of 82 Design portfolios (4.9%) with more than two standard deviations difference between results obtained from the Comparative Pairs judgements and Analytical marking. In Visual Arts the number was also quite low with only five such artworks out of 75 (6.7%), despite there was some degree of similarity between differences. In Design, a closer examination on these portfolios suggested that the discrepancies could have been caused by assessors' personal preference, portfolio qualities, and interaction between these factors and procedural factors from the scoring processes. In Visual Arts, the interplay between the qualities of the artworks, assessors' personal preference, and the procedural factors from the scoring processes seemed to be more intricate than in Design. While both tasks involved numerous components such as originality, skills, techniques, and others, in Design these components were more visible even though they were spread throughout the 15 pages. In contrast, in Visual Arts the evidence of the presence of these components could be obscure within the artwork and the idea behind it. As a result, for certain portfolios or artworks the scoring result from the Comparative Pairs judgements could be very different to the result from the Analytical marking even though the criterion for the Comparative Pairs judgements was based on the criteria for the Analytical marking.

Comparative Pairs misfits

For most Design portfolios with a large weighted mean score (wms) difference to the mean value, assessors' notes from the ACJ system indicated that these portfolios had contradictory qualities which made judgements difficult. These notes further supported the inference that in Design assessors often had the propensity to value either *process* or *product* more than the other. When a holistic criterion was used to judge these kinds of portfolios, this tendency could become more prominent because assessors needed to balance these contradictory qualities that may be in a detailed marking rubric. There was indication that it was quite similar for the Visual Arts portfolios but the notes were less conclusive. In general there seemed to be hesitation among Visual Arts assessors on the overall quality of the artworks relative to the artworks to which they were being compared. This could have been caused by incomplete submissions or confusing artwork components. It was interesting that even though there was hesitation, the quality of the digital representation in Visual Arts were not mentioned as a problem.

Validity of assessment

An analysis of the validity of the assessment of digital representations of creative work using the Comparative Pairs judgements method was based on the reliability of resulting scores, comparison with results from other scoring methods, and validity issues emerging from the assessor interviews. In general these three points of reference suggested that the Comparative Pairs judgements, as implemented in this study, could be a sufficient assessment method for the types of practical tasks in the Design and Visual Arts courses.

Analysis of the scores obtained from the two online scoring methods in both Design and Visual Arts indicated good internal reliability. The reliability coefficients of the scores obtained from the Analytical marking were 0.962 in Design (Table 4.19) and 0.944 in Visual Arts (Table 5.20). These coefficients represented the internal reliability of the scores. The reliability coefficients generated by the ACJ system for the Comparative Pairs judgements were 0.941 in Design and 0.959 in Visual Arts. As was discussed before, while the reliability coefficients generated by the ACJ system represented both the internal reliability of the scores and the inter-rater reliability of the assessors, the reliability coefficients in the Analytical marking only calculated the internal reliability. The inter-rater reliability in Analytical marking was represented by the correlations between the scores from the Analytical marking assessors. While the internal reliability from the scores from both the Analytical marking and the Comparative Pairs judgements was similarly high in the two courses, the correlations between Analytical marking assessors were only low to moderate. This suggested that there was only low-to-moderate agreement between these assessors. Correlations between scoring methods in both Design and Visual Arts indicated that the Comparative Pairs judgements results significantly and moderately correlated with results from the other two scoring methods.

Issues that could affect the validity of this form of assessment, as was reported by assessors in the two courses, were quite similar. The main concerns they had were from the quality of the digital representations and potential problems regarding subjectivity. In Design even though the assessors reported several problems concerning the quality of the digital portfolios, they considered those problems to be minor and did not overly affect the quality of their judgements, especially because in Comparative Pairs judgements the judgements were less detailed than in Analytical marking and involved only comparing two works. Visual Arts assessors were more concerned about the quality of the digital representations than the Design assessors, especially because details such as textures, dimension, and colours were either indiscernible or appeared different to the original. Most of them reported the quality of the digital representations might affect the accuracy of their judgements.

In Design, several assessors also reported the inconvenience of navigating through the digital portfolios but neither of them considered this to affect their judgements. They found it easier to flick through the original paper portfolios, especially when they needed to focus on details, than zooming in and out on the PDF files. However, similar to their perception on the quality of the digital representations, because in the Comparative Pairs judgements they only needed to compare a pair of portfolios, they did not consider this inconvenience to affect their judgements. In Visual Arts there was no such report, which was likely because the digital representation was of a finished artwork presented in files containing single images (also collated into a PDF) and a short video. Visual Arts assessors, however, found the inability to view the paired works side-by-side to be challenging. Regarding the online assessment systems, there was no problem reported in either course. Both Design and Visual Arts assessors considered the systems were easy to use.

Concerns regarding the Comparative Pairs judgements in both Design and Visual Arts included assessors' hesitation when they compared two different types of work such as photography versus technical graphic in Design and two-dimensional versus three-dimensional artworks in Visual Arts. A judgement was also considered difficult when the two works being compared were of similar quality. This tended to occur later in the Comparative Pairs judgements, when the system paired works with increasingly more similar qualities.

Design assessors tended to be more accepting towards using the Comparative Pairs judgements method with the method considered to be more accurate, straightforward, and objective; and there were no major problems related to the digital representations.

They also agreed that results from the Comparative Pairs judgements would be more reliable than the Analytical marking because it involved many assessors. On the other hand, while the Visual Arts assessors reported that they did not consider the digital representations to be suitable for the course, they considered the Comparative Pairs judgements to be more suitable than the Analytical marking.

Discussion Addressing the Research Question

Analysis of data in this study was aimed to address the overarching research question:

> *How representative are the Comparative Pairs judgement scores of the quality of the student practical production work in Visual Arts and Design courses?*

This section will discuss the findings from the study in terms of the three subsidiary research questions:

> *In assessing student practical work in each of the Visual Arts and Design courses,*

- *How valid and reliable are the scores and rankings generated by the Comparative Pairs judgements?*
- *What are the differences and similarities of the results from the Comparative Pairs judgements with the traditional analytical marking?*
- *How do the different types of work in Design and Visual Arts courses affect the scores and rankings generated by the Comparative Pairs judgements?*

Validity of assessment and reliability of scores

The first subsidiary question was:

> *In assessing student practical work in each of the Visual Arts and Design courses how valid and reliable are the scores and rankings generated by the Comparative Pairs judgements?*

The validity of the Comparative Pairs judgements were analysed using a validation framework developed by Kane (2006), Shaw et al. (2012). This validation framework was based on an analysis of the evidence for validity and threats against validity in five validity inferences (*Figure 2.5*). Hence, validation was seen as an evidence-gathering exercise. In this study, only the two first inferences were considered relevant: *construct representation* and *scorin*g, because this study mainly only examined the scoring results.

Construct representation

With regards to the *construct representation* inference, as presented in Shaw et al.'s validation framework (2012, p. 167), the first measure was the general definition of validity, which was the extent to which an instrument measures the constructs it aims to measure (Frisbie, 1988). As such, the first indication that the assessment tasks measured the constructs the courses aimed for the students to achieve was the alignment between the syllabi, the assessment tasks, and the criteria. The assessment tasks investigated in this study were the WACE examination practical component developed by the Curriculum Council of Western Australia, as expounded in the requirement documents (Curriculum Council of Western Australia, 2011b, 2011c).

The course content in the Design course was built from three content areas: design principles and process, communication principles and visual literacies, and production knowledge and skills, as was disclosed in the Design syllabus (Curriculum Council of Western Australia, 2010a). The Design course outcomes included design understandings, design process, application of design, and design in society. The practical assessment task in the Design course was a portfolio that exemplified the students' design process that led to finished design projects. The holistic criterion used in the Comparative Pairs judgements method for Design was *Judgement about performance addresses students' ability to apply elements and principles of design in recognising, analysing and solving specified design problems innovatively with consideration for a target audience and justify design decisions through experimentation and production.* An inspection of the key terms and general understanding of the syllabus, task, and criterion indicated that the three assessment components were aligned, indicating that the task and criterion should measure the intended outcomes.

In Visual Arts, there were two course elements which were art making and art interpretation (Curriculum Council of Western Australia, 2010b). Within art making, the components were inquiry; visual language; visual influence; art forms, media and techniques; art practice, presentation, and reflection. The components within art interpretation were visual analysis; personal response; meaning and purpose; and social, cultural and historical contexts. The assessment task for the Visual Arts course was a finished artwork, with intended outcomes: creativity and innovation, communication of ideas, use of visual language, use of media and/or materials, and skills and/or processes (Curriculum Council of Western Australia, 2011c, p. 5). The holistic criterion used in the Comparative Pairs judgements method for Design was *Judgement about performance addresses students' ability to creatively use visual language, materials and processes to skilfully communicate an innovative idea in a resolved artwork*. Similar to the Design course, the terms and general understanding of the three assessment components were aligned, indicating that the task and criterion should measure the intended outcomes.

Another issue related to the construct representations was the possible difference between the scoring criteria and the assessors' varied understandings of the criteria, or as Pollitt (cited in Kimbell et al., 2009) articulated "differences in conceptualisation of the trait being measured" (p. 79). This difference could be a source of threat to construct validity and score reliability. The discrepancy analysis that was conducted on student works that were scored quite differently in the two online scoring processes and on student works that the ACJ system indicated as misfits (i.e., works that had inconsistent judgements in the Comparative Pairs judgements) indicated that assessors' bias could be a source of this difference.

In the Design course, there was evidence that assessors had an inclination to value either *process* or *product* more than the other. *Product* could refer to either the observed quality of the design products or the quality of the portfolio including the visual quality and portfolio management. The Design task was a collection of evidence of a Design process. While the holistic criterion specified *process* qualities with keywords such as *recognising*, *analysing* and *solving*, *consideration for a target audiences*, and *justify design decisions* *through experimentation and production*, notes from the ACJ system indicated that particularly in close judgements assessors could value *process* more with comments such as *Neither strong but B more original* and *A shows more innovation* while others could value *product* more with comments such as *close judgement, A better resolved* and *a stronger design aesthetic is shown in B*. In Design the students had to showcase their Design process, and the holistic criterion represented that requirement. However, when the quality of the portfolios was similar, assessors' personal propensity could be the deciding factor.

In Visual Arts, the assessors' preferences were more varied with assessors leaning towards one or several qualities more than others; for example, qualities such as technique, skill, finish, idea, originality, and innovation. This difference between the two courses might be due to the difference in the type of task. The Visual Arts task was a finished artwork accompanied with an artist statement and an installation picture. Assessors judged the student work based on pictures and videos of the artworks. The finished artworks could show several qualities stronger than the others. The qualities emphasised in the holistic criterion included *creatively use visual language, materials and process*, *skilful communication*, and *innovative idea*. Unlike in Design, Visual Arts assessors' preferred qualities were more varied, especially when the judgements were close. Such comments were: "B has communicated more effectively whilst A has better skills" in which B was the winner, quite possibly because this assessor was more drawn towards visual communication, "B has more evidence but A is more cohesive" in which A was the winner, and "I prefer the concepts and the approaches of A but the skills of B" in which B was the winner.

In both Design and Visual Arts courses there was agreement among the assessment components which were the syllabi, the tasks, and the criteria. This could be regarded as evidence for the *construct representation* validity inference. The tasks and criteria measured the outcomes intended in the syllabi. On the other hand, there was indication that assessor bias could be a source of threat for this validity inference. Assessors' notes on the ACJ system on student works that were scored differently in both courses indicated that assessors' personal criteria could be the deciding factor in the Comparative Pairs judgements, especially when the paired works were similar.

Scoring

The second validity inference from Shaw et al.'s validation framework (2012, p. 167) is *scoring*. The reliability of scores obtained from a scoring process indicates the accuracy of the scores and the consistency of the assessment, and thus validity (Anastasi & Urbina, 1997). It represents how likely it is that the variability of the test scores was due on chance as opposed to systematic errors. Reliability signifies the confidence with which decisions for students could be made, based on test results (Frisbie, 1988). Two reliability estimates commonly used to assess the reliability of test scores were internal reliability, which estimates how consistent the test scores are, and inter-rater reliability, which estimates consistency among assessors.

In this study, the reliability coefficient calculated by the ACJ system used in Comparative Pairs judgements represented both the internal reliability of the scores and the inter-rater reliability among assessors. Because of the way the judgement system was designed, judgements could be concluded when the reliability level was sufficiently high, except in the case of too many extreme judgements. The reliability of scores obtained from the Comparative Pairs judgements showed high reliability for both Design and Visual Arts. The reliability of scores contributed to the validation of an assessment (Cronbach, 1971; Pollitt, 2012c), therefore the reliability of scores obtained from the Comparative Pairs judgements in this study was also used to consider the validity of the assessment process. While reliability does not necessarily indicate validity and the attempt to improve reliability could even reduce validity (Kane, 2006), the reliability in this study was considered to be a measure of validity because the reliability coefficients represented both the internal reliability and the inter-rater reliability of the 10 Design assessors and 15 Visual Arts assessors. Agreement among such a number of experienced assessors in itself could be argued to be an evidence for validity.

Beside the reliability of scores, information regarding the validity of the Comparative Pairs judgements was also obtained from comparisons between this method and other methods, namely the Analytical marking and the official WACE marking. These comparisons indicated that the scores from the Comparative Pairs judgements in each course were significantly and moderately correlated with scores from the other two scoring processes. This further showed that there was consistency between scores obtained from the Comparative Pairs judgements method and the other two methods (Shaw et al., 2012).

Factors that could reduce the validity of the Comparative Pairs judgements were inferred from the assessor interview. In general three major factors in both Design and Visual Arts courses were identified. These were the quality of the digital representations, the assessors' uncertainty when they had to compare two different types of work, and the difficulty of judging two works that had similar qualities. Design assessors reported that while there were several digital portfolios that were not clear, for example when the students used pencil to write or draw, they did not encounter any substantial problems related to the quality of the PDF file. Therefore, the quality of the digital representations was not considered as a threat to validity in the Design course. On the other hand, all Visual Arts assessors considered the quality of the photographs and videos of students' artworks to be low hence could reduce the validity of their judgements. In both courses several assessors reported their concern over the accuracy of their judgements when they were presented with two works that were of similar quality or of different types, for example between a portfolio in Technical Graphics and Photography in Design or between a two-dimensional and a three-dimensional artworks in Visual Arts.

Comparative Pairs judgements and Analytical marking

The second subsidiary question was:

> *In assessing student practical work in each of the Visual Arts and Design courses what are the differences and similarities of the results from the Comparative Pairs judgements with the traditional analytical marking?*

Concerns regarding the reliability and validity of assessment on creative production drove the effort to find assessment methods that could best measure student achievement. The current assessment method is the Analytical marking method, in which a detailed analytical marking rubric is used. Analytical marking rubric serves as a marking guideline which should increase the reliability of the scores and the validity of the judgement (Jonsson & Svingby, 2007). Rubrics contained scoring criteria, descriptors, and rating scale that help assessors assign scores that should well represent student achievement.

The Comparative Pairs judgements method is another scoring method that is procedurally different to the Analytical marking method but also has the potential to generate reliable scores, especially for subjective tasks (Pollitt, 2004; Thurstone, 1927). This method was first introduced in the 1920s but was recently made more feasible by the advancement in computer technology (Bartholomew & Connolly, 2017). Instead of assigning a score to each criterion such as in the Analytical marking, assessors judged the winner between each pair of work based on a criterion or a set of criteria. The advantage of this method is the better ease of judgements than in the Analytical marking, the number of judgements and assessors, which could make the result potentially more subjective, and the information that could enrich the scoring data quantitatively and qualitatively.

Findings from this study indicated that in general the scoring data from these two scoring methods were highly comparable. In both processes there were very few student works that obtained scores that were too different from the assessors. This study found that even though there were procedural differences between the Comparative Pairs judgements and the Analytical marking, both scoring methods could be used to score creative production tasks well.

The most noticeable difference between the scores from the Comparative Pairs judgements and the Analytical marking was the agreement among assessors. Results from the Comparative Pairs judgements indicated consistency in assessment judgements, as indicated by the inter-rater reliability, while in the Analytical marking, the comparisons of results among assessors indicated inconsistency, as shown by the correlation coefficients between assessors in each course. This indicated that even with the use of a detailed marking rubric, assessors’ personal judgements could still vary. These personal judgements could be influenced by factors such as assessor experience, assessor preference, and visual appearance of the piece of work. While these factors could still affect the judgements in the Comparative Pairs process, the effect was less substantive because of the number of judgements and assessors that this process warranted. Even though the correlation coefficients between Analytical marking assessors indicated some degree of inconsistency, there were only a few portfolios that were scored very differently in both methods, indicating that the inconsistency was spread among the scores.

Comments from the assessor interviews indicated the Comparative Pairs judgements process was considered easier than the Analytical marking process. The scoring time from the two online processes also indicated that the Comparative Pairs judgements process in each course took either similar or even less time than the Analytical marking process. The Comparative Pairs judgements were considered easy because the holistic criterion was used and there was no process of matching components on student work with the descriptors on the marking rubric. Assessors from both courses reported that the holistic criterion was easier to remember and the comparing process was faster because it only required them to judge the winner of two works. Even though the two courses had different types of tasks, the assessors’ comments from the two courses and the scoring time indicated that the Comparative Pairs judgements method was more convenient than the Analytical marking, regardless of this difference.

Concerns regarding the accountability of the results of the Comparative Pairs judgements method were reported by assessors from both Design and Visual Arts. The use of analytical marking method with a set of criteria has been largely accepted and used in education assessment (Madaus & O'Dwyer, 1999), hence an alternative method that was simpler could quite understandably cause concerns. In general, assessors in both courses expressed their hesitation over the accuracy of their judgements. Additionally, there was also apprehension over the reaction of the assessment stakeholders, especially parents and students, on the use of a single statement as a marking key.

Assessors' comments from the interview and their notes on the ACJ system indicated that in Comparative Pairs judgements method, the visual appearance of the portfolios and artworks could substantially influenced judgements, especially when the works being compared were of similar qualities. In cases when the digital representations were less representative of the original works, assessors in both courses reported that the Comparative Pairs judgements method was better than the Analytical marking. Additionally, assessors' notes on the ACJ system indicated the assessors could judge the pair based on the way the works' visual presentation such as portfolio typography and layout in Design, and attractiveness in Visual Arts. These notes indicated that this tendency was more pronounced when the works being compared were of similar quality. Since the official marking rubrics contained the elements of *visual communication* in Design and *visual language* in Visual Arts, this inclination should be a judgement factor. However, while in the Analytical marking rubric the score range could guide the assessors, in the Comparative Pairs judgements method, it was left to each assessor's judgement. Furthermore, for Design, unlike in Visual Arts which attractiveness was an integral part of the artworks, the visual literacy applied to both the design products and the presentation of the portfolio. Hence, for Design this tendency could potentially skew the judgement more than for Visual Arts.

The effect of the Assessment Tasks on judgements and scores

This section discusses the third subsidiary question:

> *In assessing student practical work in each of the Visual Arts and Design courses, how do the different types of work in Design and Visual Arts courses affect the scores and rankings generated by the Comparative Pairs judgements?*

The types of work in the two courses were very different and the manner in which they were presented as digital portfolios was very different. It was likely that these differences would affect the use of the comparative pairs method of judging and scoring. In particular this method requires a holistic judgement so there could be differences in the likelihood that such judgements could be readily made. This may be evident in the reliability of the scoring, the time taken to make judgements and the comparative validity of the judgements compared with other methods of scoring.

The assessment task for Design was a 15-page design portfolio containing evidence of a design process for up to three projects. In Visual Arts the task was a finished artwork. For Design variations in student work included the type of work, for example photography and technical graphic, the format of the portfolio, and the type of evidence used. In Visual Arts variations in student work included the type of artwork such as two-dimensional and three-dimensional, the media of the artwork, and the dimension of the artwork. As such, there were different challenges for the two courses that could influence the suitability of the Comparative Pairs judgements. In general, the Design task was a *process* and in Visual Arts it was a *product*.

Analysis of findings suggested that in general the Comparative Pairs judgements could be suitable for both types of task. The reliability of the scores obtained from the two courses was high, there were relatively very few works that were judged too differently, there were relatively very few inconsistent judgements among assessors, and even though there were challenges reported in both the notes from the ACJ system and comments from the assessor interview, those challenges did not seem to be regarded as a cause for concern. Furthermore, assessors in both courses found this scoring method to be easier than the Analytical marking method and the scoring time indicated that the time needed in the two courses were similar, even though the tasks were different.

Comparisons with other scoring methods indicated moderate and significant correlations with both the Analytical marking and the WACE official marking results in the two courses. In the Comparative Pairs judgements process holistic criteria were used on digital representation of student work, while in the Analytical marking it was analytical rubrics on digital representation and in the WACE marking it was analytical rubrics on original work. Good correlations with the other two methods consequently suggested that these procedural differences did not strongly affect differences of results from the three methods. Discrepancy analysis in the two courses indicated that assessors’ tendency towards particular criteria could be a problem in Comparative Pairs judgements. In both courses, this tendency seemed to be more prominent when the overall quality of the two works being compared was similar.

For Design there were mainly two general criteria towards which assessors tended to lean, which were *process* and *product*. While several assessors tended to look for evidence of *process*, there were others who looked more for evidence of *product*. This could be due to the principles in Design and Design education. Huygen (1997) asserted that "Design and the applied arts remain an area that vacillates between artistic and economic practice, between ideals and their realization" (p. 41), suggesting that the principles of Design are driven by conflicting yet integrated factors that are theoretical or philosophical, and practical. When pedagogy is added to the mix, i.e., in Design education, these factors become even more complicated. Stables (2017a) further explicated different issues within Design education which included social issues such as consumerism and ecological issues such as sustainability. These conflicting principles of Design education made it plausible, thus, that when Design assessors were presented with works of similar quality, they drew upon their personal Design principles and judged the pair based on either process or product. Assessors who considered the theoretical and philosophical Design principles important might lean more towards process. Accordingly, assessors who were more practical might lean more towards product.

For Visual Arts the propensity was more varied, with assessors valuing certain qualities more than others such as originality, skills, and technique. These variations might be due to the nature of Visual Arts and the nature of the Visual Arts task in this study. Visual Arts encompasses a broad variety of genres, forms, materials, and processes; all in which creativity is prominent and influenced by many factors. Consequently, judgements in Visual Arts are bound to vary and subjective (Beattie, 1997; Laming, 2011; Rayment, 1999).

Comparative Pairs judgements for Design and Visual Arts tasks

Finally this discussion summarises the findings in terms of the overarching research question:

> *How representative are the Comparative Pairs judgement scores of the quality of the student practical production work in Visual Arts and Design courses?*

Findings from this study suggested that in general the Comparative Pairs judgements method could be suitable for both types of task used in the Design and Visual Arts courses. Even though the type of the task in the two courses was different, the similarities in the quality of the scoring results in both courses were quite strong. In both courses, the reliability of the scores was high, there was good comparability with other methods regardless of the procedural factors, and there were no apparent issues arising from the scoring process, scoring media, or the holistic criterion for each.

However, the structure of the Design portfolio made it difficult for assessors to make holistic judgements and this was reflected in lesser comparability with scores from the analytical marking. It is likely that the scores from the comparative pairs judgements were less valid although if some changes were made to the form of the portfolio this may be improved. By comparison assessors found it relatively easy to make judgements of the Visual Arts work rather than trying to make absolute judgements through the analytical marking. For both courses it appeared that some assessors made judgements on a dichotomous view of the holistic criterion; that is, on the basis of process-product (e.g., Design) or skills-meaning (e.g., Visual Arts).

Summary

This chapter has presented findings from the previous two chapters, Design and Visual Arts, in a cross-case analysis. This analysis was discussed in two parts, the first being from the point of view of the assessment task, and the second, the task assessment. The assessment task section compared findings from the two courses related to the nature of the student work, the digitization process, and the technical limitations. The task assessment section compared findings from the two courses related to the scoring results and the validity of the assessment. A discussion based on the research question and the subsidiary research questions followed.

Regarding the assessment task, the combination of the nature of student work, constraints from the digitisation process, and the technical limitation was considered to potentially affect the quality of the assessment in different ways. In Design, the digital representations of the student work were reported to be sufficient for the Comparative Pairs judgements process, despite similar problems encountered during digitisation and similar technical limitations. Because the Design task was a paper portfolio, the student works mostly could be scanned easily to create relatively clear digital portfolios. In Visual Arts, on the other hand, the student work was in the form of artworks that could be in various dimension, materials, and parts. Consequently, in Visual Arts the digital representations were photographs and short videos of the artworks. The nature of work at Visual Arts, combined with constraints from the digitisation process and technical limitations, created digital representations that the assessors found to be lacking in clarity.

Findings on task assessment indicated that there could be strong school cultures that influenced student achievement in both courses. Comparisons of results between scoring methods showed good correlations between the Comparative Pairs judgements with the other two methods, with discrepancy analysis indicating the possibility of assessors' bias towards either *process* or *product* in Design and towards a variety of components (e.g., inventiveness and skills) in Visual Arts. The rest of the analysis from the task assessment part was directly related to the research question; therefore it would be incorporated into the summary from the discussion on the research question.

The first subsidiary research question was concerning the reliability and validity of the Comparative Pairs judgements. The ACJ system that was used to manage the Comparative Pairs judgements calculated the reliability coefficient for each round of judgements. This correlation coefficient represents both the internal reliability of the judgements and the inter-rater reliability among assessors. In both Design and Visual Arts course this coefficient was sufficiently high. The validity of the assessment was analysed based on a validation framework using inferences to collect evidence for validity and issues that could be a threat to validity. Findings related to this validation included reliability, comparability with the other scoring methods, and issues arising from the assessor interview and assessors' notes recorded on the ACJ system.

The second subsidiary research question examined the differences between the Comparative Pairs judgements and the Analytical marking in Design and Visual Arts. Findings related this subsidiary research questions indicated that the procedural differences between the two scoring methods did not cause differences in the scoring results. However, the consistency among assessors was higher in the Comparative Pairs judgements than in the Analytical marking, in both courses. This could be due to the different number of assessors and the different processes between the two methods.

The third subsidiary research question was related to the suitability of the Comparative Pairs judgements for the different types of task in Design and Visual Arts. Despite the difference between the types of task in Design and Visual Arts, most assessors in each course reported the Comparative Pairs judgements to be easy and suitable for the tasks. However, there were expressed concerns over the fairness of the judgements, especially when the paired works were of similar quality. Results from scoring also indicated that the Comparative Pairs judgements could be suitable for the two types of task.

With regards to the main research question, the Comparative Pairs judgements method was found to be a suitable scoring method for Design and Visual Arts courses. The method showed good reliability and validity, with only a few inconsistencies and misfits. Course-specific issues were found, such as digitisation issues in Visual Arts, portfolio navigation issues in Design, and different tendencies of assessor bias in each course.

The following chapter discusses the conclusions drawn from the findings, limitations of this study, implications, and recommendations for future research.

CHAPTER 7
CONCLUSIONS

This chapter presents the conclusions from the study as implicated in the research question, followed with acknowledgement of the limitations of this study. The implications of the findings for policy and practice are discussed next, as well as recommendations for future study.

The aim of this study was to investigate the suitability of the Comparative Pairs judgements method in assessing student practical production work for the purpose of summative assessment. The validity of the assessment method was viewed as a base for suitability, with three points of reference: the reliability of the scores; comparisons with results from other scoring methods; and issues that might influence validity. This aim was built into the overarching research question for this study, which was:

How appropriate is the Comparative Pairs method of judgement for assessing the quality of student practical production work, represented in digital forms, in the Visual Arts and Design courses in Western Australia?

The two key concepts that were built into this question were *digital representations of practical production work*, which refers to the matters surrounding the digital representations especially from the task assessment perspective, and *the quality of the Comparative Pairs scoring method in assessing creative work,* which discusses the task assessment. Conclusions on these key concepts are discussed next, based on the discussion of the research findings in the previous chapter. The discussion on the conclusions in this chapter is as described in the conceptual framework (*Figure 2.7*). Conclusions on the first key concept derived from findings related to the assessment task component of the framework, while conclusions on the second key concept were based on findings from the task assessment component of the framework.

Digital Representations of Practical Production Work

There has been growing interest in the use of digital technologies to support assessment of student learning in schools (Barber, King, & Buchanan, 2015; Griffin et al., 2012; Kimbell, 2008; Newhouse et al., 2011; Stacey & Wiliam, 2012; Timmis et al., 2016); sometimes referred to as e-assessment. This interest has both made the development of e-assessment procedures important and opened up new possibilities in the practice of authentic assessment. Current e-assessment practice varies from a transfer from pen-and-paper into computer-based assessment to Computerised Adaptive Testing (CAT) which adapts the difficulty level of the assessment to individual student's performance (Gershon, 2005).

The use of digital technologies in authentic assessment enables the flexibility of creating assessments that capture any kind of student work, assess the work in various ways, collaborate with students or other teachers, and create reports or feedback with ease. Digital capture of student work allows the recording of far more than pen-and-paper kind of student work. Current technology makes capturing performances in drama, dance, artworks, sports, and many others, feasible (Dillon & Brown, 2006; Drijvers et al., 2016; Heldsinger & Humphry, 2010; Jones & Alcock, 2012; Newhouse, 2011b; Newhouse et al., 2011). More research is needed to explore these possibilities.

Matters surrounding the digital representations of practical production work in this study are discussed based on the *types of digital representations and digitisation process,* the *quality of digital representations, and the scoring system accessibility*. These matters were considered to be the technical issues that could potentially influence the way the quality of the Comparative Pairs judgements was perceived.

Types of digital representations

Digital artefacts of student works could include digital creation (i.e., the students create the works digitally), and digital representation (i.e., students' original works were digitised). One of the most preferred type of digital artefacts is digital portfolio (Joint Information Systems Committee [JISC], 2008; Masters, 2013), because it can provide information on student progress by collating different types of tasks and feedback. In the present study, the second kind of digital artefacts was investigated, which is the digital representations of student work. In both courses, these digital representations were in the form of digital portfolios. In the Design course the digital representation was in the form of PDF files created by scanning students' original paper portfolios. In the Visual Arts course digital representation was in the form of photographs, video recording of the artworks and a PDF file created from these photographs and the accompanying artist statement. JISC defined a digital portfolio as "a collection of digital artefacts articulating experiences, achievements and learning" (p. 6), while Masters (2013) identified the benefit of using portfolios as "When assembled over a period of time, portfolios can provide a valid basis for establishing current levels of achievement and for monitoring progress over time" (p. 38). In this assertion, the use of portfolios to display the learning process is similar to the Design portfolio in the present study. In the Visual Arts course, the portfolios were more of a tool to display components and details of the students' arts products.

In Design the digital portfolio was simply the digital form of the original paper portfolio, which was the assessment task in the Design course. There was good alignment between this assessment task to both the syllabus and the criterion, which indicated that this type of presentation of student work could be considered suitable for the task. In Visual Arts the digital portfolio contained photographs of student artwork that aimed to represent the qualities of the artwork; such as creativity, innovation, and visual communication; to provide the assessors with sufficient information to make their judgements. The artwork as the assessment task was found to be aligned well with the Visual Arts syllabus and the scoring criteria. It could then be established that the digital portfolio could be a suitable form of digital representation for the Visual Arts course.

The types of the digital representations in Design and Visual Arts were considered to be a factor that influenced assessors' judgements in the Comparative Pairs scoring. In the Design course, the assessors decided on the winner of the paired works presented by comparing a variety of evidence spread around the 15-page portfolios based on the holistic criterion. In Visual Arts they tended to compare a pair of photographs of finished artworks based on the holistic criterion, assisted by other supporting photographs, video(s) and artist statement. As such, the judgement mechanism in Design was potentially more complicated than Visual Arts, however, as the Design assessors reported, this mechanism was less complicated than in the Analytical marking. In both courses, when the two portfolios being compared were of similar quality or of different course contexts, however, the Analytical marking was considered easier and more reliable.

Quality of the digital representations

Regarding the quality of the digital representations, the main problems for both courses were the clarity of the representations and discrepancies between the original work and the representation; termed the *fidelity* of the representation. On the fidelity of the representation of student work, Dillon and Brown (2006) warned "… the experience of making and perceiving is difficult to capture, and once captured may not be a true representation of the work" (p. 422). In Design, the quality mainly depended on the type of the paper and the clarity of the writing or drawing. Portfolios that were created on glossy paper, for example, were difficult to scan to create digital portfolios that represent the quality of the original. The fidelity of the Design portfolios was largely quite high, as was also indicated by the assessor interviews. Even though there were several portfolios with parts that were not clear, all Design assessors expressed their satisfaction over the quality of the digital portfolios.

In Visual Arts, the dimension, components, details, and the type of the media of the artworks affected the quality of the digital representations. As a result, there were photographs that did not sufficiently capture the colour or the details of the original artworks. Aside from that, Visual Arts assessors also reported the loss of the sense of perspective and the uncertainty of certain qualities of the artworks related to it. Regarding ways to improve the quality of digital representations, Dillon and Brown (2006) suggested:

> An effective ePortfolio will take into account this framing and flattening effects of the digital representations, either by supplementing them with additional

> material to compensate and/or by deliberately highlighting the fact that data in these forms is reductive so that the viewer takes that into consideration. (p. 422)

Included in the artwork submission requirement for the Visual Arts course were an artist statement and an installation photograph. These components were intended to provide information for assessors regarding the features of the artwork that the student chose to emphasise, such as the background or the reason the particular medium was selected. These components were also included in the PDF and PowerPoint files. Additional materials that were expected to compensate for the lack of the sense of perspective were four close-ups, a short video, and a VR video for small to medium three-dimensional works. However, the interview with the assessors indicated that this effort was less than sufficient and they expressed their concerns over the reliability of their judgements.

Conclusions from the digital representations of practical production work

Digital portfolios could be considered to be an appropriate learning artefact to represent progress of student's learning or to showcase achievements. In Design the digital portfolio was used to represent progress while in Visual Arts it was the latter. In the present study, these types of digital representations were found to be manageable and appropriate.

In the Design course, the quality of the digital representations was satisfactory; with assessors from both Analytical marking and Comparative Pairs judgements not reporting substantial problems. With PDF files that had unclear parts, for example pencil drawings that the scanning failed to capture, the Comparative Pairs judgements process was especially considered easier because it required consideration of relatively less details than the Analytical marking. The type of the digital representations in the Visual Arts course could be considered suitable for the assessment task, however, the quality of the digital presentations was not found to be satisfactory for some portfolios. All Visual Arts assessors reported difficulty in feeling confident with their judgements because they found the photographs did not quite capture the quality of the original artworks.

In general, even though factors related to the digital representations in Visual Arts might affect assessors' judgements, and consequently students' scores, this study revealed that the influence was not substantial. Comparisons between the three scoring methods did not indicate that the difference in the assessment media created differences in the scores resulting from the use of these methods in either of the courses. This suggested that while assessing the representations might be different to assessing the original works, the quality was adequate to allow experienced assessors to make adequately accurate judgements.

The Quality of the Comparative Pairs Scoring Method in Assessing Creative Work

The quality of judgements for the Comparative Pairs scoring method is represented by the reliability measures of the scores and the validity of the assessment as a whole, as was described in the conceptual framework (*Figure 2.7*). The analysis of the validity was conducted on two inferences based on Shaw, Crisp and Johnson's validation framework (2012), which were *construct representation* and *scoring*. The validation analysis based on construct representation inference was built on two points of reference, which were the alignment among the syllabi, assessment task, and criteria as was discussed earlier in this chapter; and the alignment between the scoring criteria and the assessors' interpretation of the criteria. The validation analysis

The validation analysis was based on three points of reference deriving from several validation theories (Cizek et al., 2011; Cronbach, 1971; Elliott et al., 2007; Kane, 2006; Kimbell et al., 2009; Pollitt, 2012c; Raykov & Marcoulides, 2011; Sireci, 2007). These three points of reference were reliability of the scores, comparability with other methods, and factors that could affect validity such as the accessibility of the online scoring systems. In the following sections, the alignment between the scoring criteria and the assessors' interpretation of the criteria is discussed within the factors that could affect validity as these two points of reference were found to be related.

Reliability of scores

The scores obtained from the ACJ system for the Comparative Pairs judgements for both courses had high internal reliability and inter-rater reliability. The internal reliability estimate indicated the internal consistency of the scores, which was similar to Cronbach's alpha (Kimbell et al., 2009), while the inter-rater reliability indicated the agreement among assessors (Stemler, 2004). The misfit statistics, which was another measure for the assessors' agreement and consistency of judgements (Pollitt, 2012a), also did not indicate extreme misfits. Overall, confidence on the reliability of the scores for both courses was high.

Comparisons with other scoring methods

Comparisons with other scoring methods indicated that results from the Comparative Pairs judgements were relatively similar to results from the other two sources of scores using analytical marking methods, adding to the confidence on the quality of this scoring method (Elliott et al., 2007; Haertel, 1999). Several works were ranked differently by the different methods, and further examination on scores, ranks and assessment notes indicated that the main possible reasons for the discrepancy were the relatively small sample size, and assessors' personal tendency towards certain qualities more than others.

Factors affecting validity

In the present study, there were several factors that were identified to possibly affect the validity of the Comparative Pairs judgements results. These factors were the scoring systems accessibility, assessors' bias, and the quality of the digital representations. This section discusses the first two factors. The quality of the digital representations has been discussed earlier in this chapter.

Scoring systems accessibility

In terms of the accessibility of the digital representations and the online systems used in the scoring processes, there did not seem to be any problem related to the file size or type of the digital representations for either of the courses. All assessors reported that the online systems were easy to access and relatively fast for both the Analytical marking and the Comparative Pairs judgements processes. Aside from the need to zoom in or out and to scroll through the digital portfolio instead of flicking through the pages on the original portfolio, the Design assessors considered there was no substantial problem with the online systems. In Visual Arts, the assessors preferred to be able to have the pair of works presented side by side. Aside from this and several instances when the internet connection was slow, there were also no substantial problems with the online systems for Visual Arts.

Consequently, issues concerning online scoring systems accessibility were not considered to affect the validity of the assessment results. Both the Analytical marking and Comparative Pairs judgements systems were reported to work well and was easy to use in both courses, which further indicated that the differences between the two scoring processes were unlikely to affect the comparability of the processes. Between the courses, the only notable difference related to the scoring systems was the system inability to present the paired Visual Arts works side by side, which could make it difficult for the assessors to directly compare certain features of the artworks, hence this issue could potentially affect the reliability of the scores.

Assessor bias

The subjective nature of performance tasks understandably produces assessment results that vary among different assessors, even when a set of detailed criteria is used (Humphry & Heldsinger, 2014; Miller & Linn, 2000; Pollitt, 2004; Wiggins, 1990). Research on assessor judgements suggested that assessors' judgements do not necessarily adhere to marking criteria, but variations could involve "personal choices about sources of information, about how to combine information" (Allal, 2013, p. 31).

Put simply, performance assessments position assessors "as operating in a dual mode, (a) as custodians of in-the-head standards, and (b) as experts in making complex comparisons" (Sadler, 1986, p. 8). In both the Design and Visual Arts courses, the assessors operated in this dual mode. In Design they memorised the elements of the criterion and compared the various evidence presented between the paired portfolios. In Visual Arts, the process was possibly less complicated because the task was finished artworks with assessed elements such as creativity and materials quite visible for experienced assessors. The difference between this mechanism in the Analytical marking and the Comparative Pairs judgements was the comparison used. In the Analytical marking, the comparison was the criteria, making the judgements more complicated because of the number of criteria used and the score range for each mastery level. In the Comparative Pairs judgements, as one of the assessors reported, the mechanism was relatively less complicated because the comparison was another portfolio, displayed side by side. Laming (2011) established that "people are generally poor at judgements of single stimuli by themselves; relationships between stimuli are much more accurately discerned" (p. xv), indicating that comparisons between portfolios are potentially more reliable than assigning a score on individual ones.

Wiggins (2011) argued, "That is why most so-called 'criterion-referenced' tests are inadequate: The problems are contrived, and the cues artificial … what the students need is a test with more sophisticated criteria for judging performance" (p. 85). Unlike the Analytical marking which is more criteria-oriented, the Comparative Pairs judgements reduced teachers' tendency to teach-to-the-test because the scores do not only depend on the criterion, but also on the performance of the whole cohort. With teachers and students understanding the Comparative Pairs judgements process, the emphasis on aligning students' learning to criteria is, at the very least, partly shifted to encouraging the students to do their best, which is more suitable for constructivist learning.

The Visual Arts task and criterion were to some extent the reverse of the task and criterion in Design. In Design the students showcased their Design process to represent their final products, while in Visual Arts the students showcased their finished product to represent their process. In Design the final product was implicit and the process explicit. When the quality of the Design process was similar, assessors could be compelled to fall back to their personal preference and made a final decision based on whether they valued more *process* qualities or the appearance of the *product*. In Visual Arts the final product was explicit and the process implicit. When the quality of the artworks seemed similar, the deciding factor was likely to not be *process* or *product*, but the qualities of the perceived process that were represented in the product.

In general, even though the interview with the assessors in the two courses showed a degree of scepticism in the use of digital representations in the Comparative Pairs judgements, the scoring data and analysis of the assessors' notes on the digital scoring systems did not indicate that the technical differences between the traditional (WACE examination) marking and this online scoring process systematically affected the differences in the assessment outcomes.

Limitations of the Study

There were several limitations that could influence the results of this study, one being the sample size. Because this study was the first stage of the main research project and also because the two courses studied generally had a small number of students per class, the sample size in this study was relatively small. The Rasch analysis that was used in the ACJ system usually required bigger sample sizes, however, since the results of the analysis were statistically sound, there was no concern over this sample size.

Another limitation concerned the rubric for Design was that at the time of data collection, Design was a new WACE examination course, therefore the analytical marking rubric used in the Analytical marking and the WACE marking still had perceived weaknesses. The perceived weakness in the quality of the Design rubric was reflected in the findings from the two marking processes in which the rubric was used, on the comparisons with results from the Comparative Pairs judgements, and on the comparisons with findings from Visual Arts. While this might skew the findings to some extent, it also enriched the research data with findings about scoring criteria, especially when Design was compared with Visual Arts, whose rubric was more developed.

A further limitation was that being the first stage of a bigger project, there were likely to be several factors that could be avoided or improved for later stages of the main project, for example improved guidelines for ensuring the quality of the digital representations. Even though the digitisation guidelines were designed through extensive consultations with experts, only after the experience did the researchers understand the situation and challenges. Parallel to this, on the Comparative Pairs judgements process, the assessors only started to understand the concept of comparing two works instead of assigning scores on a rubric as they went through the process. This learning process was likely to have affected the results of the early rounds of judgements, and therefore affected the pairings created by the system. However, the ACJ system only found relatively few inconsistent judgements associated with these early rounds.

Implications for Policy and Practice

The growing concern over assessment quality, the decreasing confidence in the value of assessment, and the fast development of ICT have created the need for educational assessment that is authentic, practical, useful, and accountable (Timmis et al., 2016) . The direction in which educational assessment is currently going is towards digital, online assessment (JISC, 2010). Educational institutions around the world have been gradually moving their assessment practice from pen-and-paper towards online assessment. This transformation warrants the development of new research-driven assessment policy and practice.

In this study Design students created their portfolios on paper and researchers from the main project scanned these original paper portfolios. As a result, there were several portfolios that were different in quality to the original or with components that were not clear enough. When students create their own portfolios digitally; this would be less of a problem. They would learn valuable new skills most likely relevant to their future study and employment, and there would be more flexibility as computer applications have many features that were lacking in a paper portfolio. Applications such as Autodesk Revit and Adobe Acrobat enable students to design, annotate, combine, edit, and many more with ease to create good quality portfolios. On the teacher and assessor’s side, digital portfolios could easily be submitted online, scored, annotated, recorded, and sent back to students or other teachers for moderation.

Visual Arts students also should be able to create their digital portfolio of their artworks, regardless of whether their original artworks were digital or not. Visual Arts students need to be able to take pictures and videos of their artworks because they know their work best and therefore they should know how to create a compelling representation of the work. The flexibility and possibilities that digital portfolios bring to Visual Arts are even more pronounced than in Design, both logistically and economically. With the use of Visual Arts digital portfolios, the process of carefully packing up and sending artworks for marking is not necessary, neither is risking damage to the artworks in transit. This is especially significant for Western Australian remote schools. Also, such practices will develop student capability to present their work in digital forms, which is increasingly an expectation in the worlds of Design and Arts. There is increasingly an expectation that such work will be presented in digital portfolios and therefore students in schools need to develop expertise in doing this.

The digital version of student creative work, regardless of whether it was created digitally or digitised, comes with the flexibility that was not available otherwise. When student creative work is digital, it may be submitted online from anywhere and assessed by assessors anywhere. It could be made available for peer assessment and moderation. It can be stored and transferred economically without a damage risk. It may be replayed, zoomed in, slowed down, and digitally annotated easily for detailed analysis for scoring or for clearer feedback.

In both Design and Visual Arts, the results from the Comparative Pairs judgements were favourable. Comparisons with the Analytical marking indicated that the Comparative Pairs judgements method is potentially more suitable for creative practical production than the Analytical marking method. The statistical analysis and interview with assessors from the two courses suggested that this scoring process was statistically defensible. Beside that, the Comparative Pairs judgements method was considered to be easier to do than the Analytical marking method. The Comparative Pairs judgements method has been signified to be a valid and a more appropriate method of scoring creative practical work, compared to the Analytical marking method. Therefore the results of this study would suggest that assessment authority bodies consider this method for assessing creative work. The Comparative Pairs judgements could be implemented in formative assessment, especially because of its feedback feature; or for online moderation; or for summative assessment.

The implementation of the Comparative Pairs judgements in the assessment of practical production would be more valuable if the assessment task is adapted to making holistic judgements. The Design portfolio, for example, consisted of evidence of design processes for up to three projects. While this task could showcase the students' understanding and ability well, the length of the portfolio combined with the varied written and image evidence presented throughout the portfolio made the judgements difficult. Assessors had to mentally hold the criterion to be used to compare the variety of evidence. A more structured task that does not limit students' creativity could be designed to better suit holistic judgements depends on the purpose of the assessment.

Professional learning activities for teachers and assessors on assessment have been considered to be essential to improve assessment quality. Most teachers and assessors have had training and experience in constructing and using analytical marking rubric, since analytical marking has been used broadly in education. If the Comparative Pairs judgements with a holistic criterion is to be gradually implemented in educational assessment, teachers and assessors need to be engaged in professional learning activities on making holistic judgements and holding competing criteria. A professional learning activity that could be suitable for this purpose is community of practice (Wenger, 2011). Wenger defined communities of practice as “groups of people who share a concern or a passion for something they do and learn how to do it better as they interact regularly” (p. 1). Teachers and assessors could align and discuss holistic judgements based on their subjects, as well as collaborate in assessment projects.

With the current direction in educational assessment, schools need to be encouraged to implement more use of digital technologies to support assessment processes (Condie, Munro, Seagraves, & Kenesson, 2007; Joint Information Systems Committee [JISC], 2008; Masters, 2013). Particularly in courses such as Design and Visual Arts, the realisation of ICT Capability as one of the Australian National Curriculum General Capabilities should include several provisions. These provisions include necessary equipment such as scanner and digital cameras; computer applications such as Autodesk Revit, Adobe Acrobat Pro and Adobe Photoshop. To ensure the ability of teachers and students to make use of the available technology, teachers and students need to be trained. Training in this area would help teachers and students to become adept in using digital technologies for assessment, giving feedback, conducting moderation, and record keeping.

Recommendations for Future Research

The importance of the use of student learning data to improve teaching and learning is as elucidated by (Hattie, 2003): "… we should be asking where the major source of variance in student's achievement lie, and concentrate on enhancing these sources of variance to truly make the difference" (p. 1). The digital form of student work combined with the Comparative Pairs judgements results and notes can be a valuable data source to identify the factors that influence student achievement. Further research on different tasks or student demographics could be focussed to find information to improve learning, diagnose challenges, or initiate subject-related changes. A comparison between results from the Comparative Pairs judgements and the other scoring methods in this study indicated strong academic school culture that influence student achievement. A more in-depth study within school subjects could be conducted either in school level or state level to further identify the various factors that contribute to student achievement, the way these factors influence student achievement, and ways to use this information to improve student achievement.

Parallel to one of the recommendations from Timmis et al. (2016) and Joint Information Systems Committee (2010), a collaboration among educational stakeholders is important. Teachers, technology researchers, educational researchers, government bodies, and industries need to work together to find best e-assessment strategies that are aligned to pedagogy. This study employed theories from early educational measurement history (Thurstone, 1927, 1928) that have been made feasible by current development in digital technologies.

Quality and type of digital representations of student work

In authentic assessment students perform tasks that are similar to real life situations such as designing a piece of furniture that is ergonomic, or creating and performing a dance.

Because of the nature of this kind of task, digital artefacts are often the best way to represent the performance for the purposes of assessment (JISC, 2010; Masters, 2013). However, as was found in this study, the quality of the digital representations of student task should have adequate fidelity to ensure the quality of the results of the assessment. Future research is needed to examine different types of tasks, and the best way to create digital representations of the tasks as a way to ensure the quality of the assessment results.

Suitability of the Comparative Pairs judgements in other types of work

The concept and use of authentic assessment have been around for decades, but the availability of information regarding the design and quality of authentic assessment is still quite limited (Scardamalia et al., 2012; Stobart, 2010). Further, quality research and practice on the use of digital technologies to support authentic assessment is even more needed (Lim et al., 2013; Masters, 2013). Aligned to the recommendations for policy and practice, further research is needed to better understand issues surrounding digital assessment in general, and specifically the use of Comparative Pairs judgements method in specific. Findings on the present study suggested that the Comparative Pairs judgements could be a suitable assessment method for practical production tasks in Design and Visual Arts which were similarly subjective and creative but also different in their nature. Considering this scoring method was also found to be valuable in other subjects such as mathematics, English, and engineering (Humphry & McGrane, 2015; Jones et al., 2015), research on different tasks within different subjects could contribute to the understanding of how the Comparative Pairs judgements could improve current assessment practice.

Research on professional learning

One of the main problems on which this study was based was the problem in assessing subjective tasks. This problem is exacerbated in subjective tasks that involve creativity and innovation, for example an artwork or a science research project. Other studies, as well as this one, consider personal preference or assessor bias to be a main factor that influences assessors' judgements (Allal, 2013; Bloxham et al., 2016; Sadler, 1986), even when an analytical marking rubric is used (Pollitt, 2004; Wiggins, 2011). Assessors' personal preference could affect the reliability of the scores, which in turn would affect the fairness of the assessment. In high-stakes assessment in particular, all efforts should be taken to ensure that the result of the assessment is accountable. Professional learning is considered to be an effective method to align assessors' judgements. Currently in Western Australia, this kind of professional learning could be in the form of marking moderation, assessor briefing, and the use of exemplars . However, a focussed research on the evaluation of such programs in tasks that are subjective and involve creativity and innovation is still limited.

Comparative Pairs judgements method for other types of assessment

Assessors notes in the ACJ system could be utilised to provide feedback on how each student could improve their work, based on the work of their peers aside from the prescribed outcomes. The advantage of this kind of feedback is it allows for achievement beyond the outcomes. In creative tasks in particular, this allows for more realistic feedback because the students could compare their work to their peers' and the teachers' comments. Research on the use of the Comparative Pairs judgements in other types of assessment such as in formative assessment or diagnostic assessment could benefit both students and teachers. As (Redecker & Johannessen, 2013) emphasised, formative assessment holds an important role in teaching practice and student learning, especially because it could provide the feedback necessary to improve teaching and learning. The use of the Comparative Pairs judgements method with a holistic criterion in formative assessment would provide both the teachers and students with information on student learning from a different perspective to analytical marking. Unlike the analytical marking that was considered to hinder creativity and condition them to measure their learning to the pre-set outcomes (Kohn, 2006), feedback from the Comparative Pairs judgements would be far less restrictive.

School culture and student achievement

Findings in both Design and Visual Arts suggested a strong relationship between school culture, components that are related to teaching and students, and student achievement.

Related to these findings, the socioeconomic status (SES) of the area of a school has been identified as one of the factors that could influence student achievement (Marchant & Finch, 2016). Other studies have found that students' ICT use positively influences their academic achievement (Araya et al., 2015; Skryabin, Zhang, Liu, & Zhang, 2015), and that students' ICT competence were influenced by their SES (Aesaert et al., 2015). Lately, mobile devices have increasingly become affordable. Students can now more easily access the online resources with an inexpensive mobile phone and a broadly available WiFi connection. The digital divide created by the SES gap is becoming smaller. Research on effective ICT implementation to improve academic achievement in areas with low SES is much needed to increase equity in education. Factors such as collective academic characteristics (e.g., persistence, understanding, intelligence), specific teaching methods, technology, teaching-to-the-test approach, and the availability of school facilities could benefit or disadvantage students, especially in creative subjects.

Overall Conclusion

The direction of educational assessment is towards computer-based and online assessment. Digital and online assessment has been found to be practical, flexible, economic, and mostly feasible. The steps that educational institutions and government bodies need to take are towards building assessment policies to use ICT to prioritise the needs of students and the way to make assessment fit the purpose. In high-stakes assessment of creative production, assessment needs to be authentic, representative, equitable and accountable.

Findings from this study suggested that the Comparative Pairs judgements method could be suitable for assessing performances such as through Design and Visual Arts digital portfolios. This study indicated that the use of Comparative Pairs judgements provided reliable scores, was easy for assessors to use, and is likely to be more suitable for these types of assessments than the more commonly used Analytical marking method. However if the Comparative Pairs judgements method is to be more widely used for educational assessment, improvement in several areas is needed. First, the assessment task needs to be constructed for holistic judgements to improve the construct validity of the assessment. Second, there is a need to develop a community of learning to assist teachers and assessors to construct and use holistic judgements in assessment. Finally, school and assessment bodies' ICT infrastructure needs to be improved to better facilitate the assessment.

The technologies for creating digital representations of performances and for making them available to assessors online have developed considerably over the past decade. We are now at a point where it is feasible to use such technologies to increase the authenticity, validity and reliability of high-stakes summative assessment using online comparative pairs judgements systems. While more research and improvement in several education areas need to be conducted to indicate the scope of such an approach, it will remain for education policy makers, researchers, and practitioners to argue the case and make the decisions in the interests of student futures.

REFERENCES

ACARA. (2017). Information and Communication Technology (ICT) Capability Retrieved from https://www.australiancurriculum.edu.au/f-10-curriculum/general-capabilities/information-and-communication-technology-ict-capability/

ACARA. (n.d.). Technologies. Retrieved from https://www.australiancurriculum.edu.au/f-10-curriculum/technologies/

Adie, L. (2013). The development of teacher assessment identity through participation in online moderation. *Assessment in Education: Principles, Policy & Practice, 20*(1), 91-106.

Aesaert, K., Van Nijlen, D., Vanderlinde, R., Tondeur, J., Devlieger, I., & van Braak, J. (2015). The contribution of pupil, classroom and school level characteristics to primary school pupils' ICT competences: A performance-based approach. *Computers & Education, 87*, 55-69.

Allal, L. (2013). Teachers' professional judgement in assessment: A cognitive act and a socially situated practice. *Assessment in Education: Principles, Policy & Practice, 20*(1), 20-34.

Allison, M., & Kendrick, L. M. (2015). Toward Education 3.0: Pedagogical Affordances and Implications of Social Software and the Semantic Web. *New directions for Teaching and Learning, 2015*(144), 109-119.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th international ed.). Upper Saddle River, NJ: Prentice Hall.

Anderson, T. (2016). Theories for learning with emerging technologies. In T. Anderson (Ed.), *Emerging technologies in distance education*. Edmonton, AB: AU Press.

Andrade, H. G. (1997). Understanding rubrics. *Educational leadership, 54*(4), 14-17.

Andrade, H. G. (2005). Teaching with rubrics: The good, the bad, and the ugly. *College teaching, 53*(1), 27-31.

Andrews, D. H., & Wulfeck II, W. H. (2014). Performance assessment: Something old, something new *Handbook of research on educational communications and technology* (pp. 303-310). New York, NY: Springer.

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement, 2*, 13.

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park: Sage Publications.

Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms? *Medical care, 42*(1), I-7.

Andrich, D. (2006). *A report to the Curriculum Council of Western Australia regarding assessment for tertiary selection*. Retrieved from Perth, Western Australia:

Angoff, W. (1996). Scales, norms, and equivalent scores. In A. W. Ward, H. W. Stoker, & M. Murray-Ward (Eds.), *Educational Measurement: Theories and Applications* (Vol. 2, pp. 121).

Araya, R., Gormaz, R., Bahamondez, M., Aguirre, C., Calfucura, P., Jaure, P., & Laborda, C. (2015). ICT supported learning rises math achievement in low socio economic status schools. In G. Conole, T. Klobučar, C. Rensing, J. Konert, & E. Lavoué (Eds.), *Design for Teaching and Learning in a Networked World* (pp. 383-388). Switzerland: Springer.

Archbald, D. A., & Newmann, F. M. (1988). *Beyond Standardized Testing: Assessing Authentic Academic Achievement in the Secondary School*. Retrieved from Washington, DC:

Ashford-Rowe, K., Herrington, J., & Brown, C. (2014). Establishing the critical elements that determine authentic assessment. *Assessment & Evaluation in Higher Education, 39*(2), 205-222.

Asia Education Foundation. (n.d.). BRIDGE School Partnerships. Retrieved from http://www.asiaeducation.edu.au/programmes/school-partnerships

Assalahi, H. (2015). The philosophical foundations of educational research: A beginner's guide. *American Jounal of Educational Research, 3*(3), 6.

Australian Bureau of Statistics. (2017). Western Australia (STE) (5). Retrieved from http://stat.abs.gov.au/itt/r.jsp?RegionSummary&region=5&dataset=ABS_REGIONAL_ASGS&geoconcept=REGION&datasetASGS=ABS_REGIONAL_ASGS&datasetLGA=ABS_NRP9_LGA&regionLGA=REGION&regionASGS=REGION

Australian Curriculum Assessment & Reporting Authority. (2011). *AED2203 foundations of art education : Book of readings*. Sydney: ACARA.

Baird, J. (2007). Alternative conceptions of comparability. In P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Baird, J. A., Greatorex, J., & Bell, J. F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education: Principles, Policy & Practice, 11*(3), 331-348.

Banks, F., & Williams, J. (2013). International perspectives on technology education. *Debates in design and technology education. Debates in subject teaching*, 31-48.

Barber, W., King, S., & Buchanan, S. (2015). Problem Based Learning and Authentic Assessment in Digital Pedagogy: Embracing the Role of Collaborative Communities. *Electronic Journal of e-Learning, 13*(2), 59-67.

Bartholomew, S., & Connolly, P. E. (2017). *Adaptive comparative judgment in graphics applications in education*. Paper presented at the ASEE, Columbus, OH.

Bartholomew, S., Hartell, E., & Strimel, G. (2017). *ACJ: A Tool for International Assessment Collaboration.* Paper presented at the PATT34 Millersville University, Pennsylvania, USA 10–14 July, 2017.

Beattie, D. K. (1997). *Assessment in Art Education. Art Education in Practice Series*: ERIC.

Begg, A. (2015). *Constructivism: An overview and some implications*. Retrieved from

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5-25.

BigDayta. (n.d.). Retrieved from https://bigdayta.weebly.com/the-dayta

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining twenty-first century skills *Assessment and teaching of 21st century skills* (pp. 17-66): Springer.

Blaikie, F., Schönau, D., & Steers, J. (2003). Students' gendered experiences of high school portfolio art assessment in Canada, the Netherlands, and England. *Studies in Art Education, 44*(4), 335-349.

Blanchard, J., & Moore, T. (2010). *The digital world of young children: impact on emergent literacy: a white paper*: Pearson Foundation.

Bloxham, S., den-Outer, B., Hudson, J., & Price, M. (2016). Let’s stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria. *Assessment & Evaluation in Higher Education, 41*(3), 466-481.

Bond, L. A. (1996). Norm-and criterion-referenced testing. *Practical Assessment, Research & Evaluation, 5*(2). Retrieved from http://pareonline.net/getvn.asp?v=5&n=2

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, N.J. ; London: L. Erlbaum.

Brabham, D. C. (2013). *Crowdsourcing*. Cambridge, MA: Massachusetts Institute of Technology.

Brightpath. (2017). Brightpath. Retrieved from https://www.brightpath.com.au/

Brookhart, S. M. (1999). *The Art and Science of Classroom Assessment. The Missing Part of Pedagogy. ASHE-ERIC Higher Education Report, Volume 27, Number 1*: ERIC.

Brookhart, S. M., & Chen, F. (2014). The quality and effectiveness of descriptive rubrics. *Educational Review*(ahead-of-print), 1-26.

Brookhart, S. M., & Nitko, A. J. (2008). *Assessment and grading in classrooms*: Prentice Hall.

Brown, A. R., & Dillon, S. C. (2006). ePortfolios in Arts Learning and Assessment. *Music Education Research and Innovation, 12*(1), 7-36.

Brown, G., Bull, J., & Pendlebury, M. (1997). *Assessing student learning in higher education*. London ; New York: Routledge.

Brown, G. T., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing writing, 9*(2), 105-121.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32-42.

Bruce, D. L., & Chiu, M. M. (2015). Composing with new technology: Teacher reflections on learning digital video. *Journal of Teacher Education, 66*(3), 272-287.

Burton, K. (2006). Designing criterion-referenced asssessment. *Journal of Learning Design, 1*(2), 10.

Campbell, A. (2008). *Performance enhancement of the task assessment process through the application of an electronic performance support system.* (Doctoral dissertation), Edith Cowan University, Perth, Western Australia.

Casimaty, T., & Henderson, M. (2016). *Risky business: ICT and creativity.* Paper presented at the ACCE Conference, Brisbane.

Cavanagh, R. F., & Waugh, R. F. (2011). *Applications of rasch measurement in learning environments research* (Vol. 2): Springer Science & Business Media.

Cizek, G. J., Koons, H. K., & Rosenberg, S. L. (2011). Finding validity evidence: An analysis using the Mental Measurements Yearbook.

Clarke-Midura, J., & Dede, C. (2010). Assessment, technology, and change. *Journal of Research on Technology in Education, 42*(3), 309-328.

Cohen, L., Manion, L., & Morrison, K. (2011). *Research methods in education* (7th ed.). Abingdon, Oxon ; New York: Routledge.

Condie, R., Munro, B., Seagraves, L., & Kenesson, S. (2007). The impact of ICT in schools - a landscape review. Retrieved from http://dera.ioe.ac.uk/1627/7/becta_2007_landscapeimpactreview_report_Redacted.pdf

Creswell, J. (2008). *Educational research : Planning, conducting, and evaluating quantitative and qualitative research (Third edition. ed.)*. Upper Saddle River, NJ: Pearson/Merrill Prentice Hall.

Creswell, J. W. (2012). *Educational research : Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Essex: Pearson.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Los Angeles: SAGE Publications.

Crisp, G. (2009). *Towards authentic e-assessment tasks.* Paper presented at the EdMedia: World Conference on Educational Media and Technology.

Cronbach, L. J. (1971). Test validation. In R. L. E. Thorndike (Ed.), *Educational Measurement 2nd ed.* Washington, DC: American Council on Education.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin, 52*(4), 281.

Cross, D., Shaw, T., Hadwen, K., Cardoso, P., Slee, P., Roberts, C., . . . Barnes, A. (2016). Longitudinal impact of the Cyber Friendly Schools program on adolescents' cyberbullying behavior. *Aggressive behavior, 42*(2), 166-180.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621-694). Washington, DC: American Council on Education.

Curriculum Council of Western Australia. (2010a). Design. Retrieved from http://www.curriculum.wa.edu.au/internet/Senior_Secondary/Courses/WACE_Courses/Design

Curriculum Council of Western Australia. (2010b). Visual Arts. Retrieved from http://www.curriculum.wa.edu.au/internet/Senior_Secondary/Courses/WACE_Courses/Visual_Arts

Curriculum Council of Western Australia. (2011a). *Design practical (portfolio) stage 2 and 3 - 2011 marker booklet*. Perth, Australia.

Curriculum Council of Western Australia. (2011b). *Design stage 3 WACE examination 2011 practical (portfolio) marking key*. Perth, Australia: Curriculum Council.

Curriculum Council of Western Australia. (2011c). *Visual arts stage 3 WACE examination 2011 practical (production) marking key*. Perth, Australia.

Curriculum Council of Western Australia. (2011d). *WACE Manual*. Perth, Western Australia.

Curriculum Council of Western Australia. (2011e). *WACE Manual: General Information for Senior Secondary Schooling 2011*. Western Australia. : Curriculum Council.

Davis, D. (2008). *First we see: The national review of visual education*. Canberra, Australia: Department of Education, Science and Training.

Dawley, L., & Dede, C. (2014). Situated learning in virtual worlds and immersive simulations *Handbook of research on educational communications and technology* (pp. 723-734): Springer.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

Dermo, J. (2009). e ‐ Assessment and the student learning experience: A survey of student perceptions of e ‐ assessment. *British Journal of Educational Technology, 40*(2), 203-214.

DiCerbo, K. E., Behrens, J. T., & Barber, M. (2014). *Impacts of the digital ocean on education*. Retrieved from London: https://www.pearson.com/content/dam/one-dot-com/one-dot-com/global/Files/about-pearson/innovation/open-ideas/DigitalOcean.pdf

Dietel, R., Herman, J., & Knuth, R. (1991). What does research say about assessment. *NCREL, Oak Brook.* Retrieved from http://www.education.umd.edu/EDMS/MARCES/mdarch/pdf/msde000013.pdf

Dillon, S. C., & Brown, A. R. (2006). The art of ePortfolios: insights from the creative arts experience. *Handbook of Research on ePortfolios*, 420-433.

Dorn, C. M., Madeja, S. S., & Sabol, F. R. (2004). *Assessing expressive learning : a practical guide for teacher-directed, authentic assessment in K-12 visual arts education*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Doug, B. (2005). From fine art to visual culture: Assessment and the changing role of art education. *International Journal of Education through Art, 1*(3), 211-223.

Drijvers, P., Ball, L., Barzel, B., Heid, M. K., Cao, Y., & Maschietto, M. (2016). Uses of Technology in Lower Secondary Mathematics Education *Uses of Technology in Lower Secondary Mathematics Education* (pp. 1-34): Springer.

Driscoll, M. P. (1993). *Psychology of learning for instruction*. Boston: Allyn and Bacon.

Dunleavy, M., & Dede, C. (2014). Augmented reality teaching and learning. In J. M. Spector, D. M. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (pp. 735-745): Springer.

Edwards, S., Nolan, A., Henderson, M., Skouteris, H., Mantilla, A., Lambert, P., & Bird, J. (2016). Developing a measure to understand young children's Internet cognition and cyber-safety awareness: a pilot test. *Early years, 36*(3), 322-335.

Eisner, E. W. (2002). *The arts and the creation of mind*. New Haven: Yale University Press.

Elliott, S. N., Compton, E., & Roach, A. T. (2007). Building Validity Evidence for Scores on a State ‐ Wide Alternate Assessment: A Contrasting Groups, Multimethod Approach. *Educational measurement: Issues and practice, 26*(2), 30-43.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*: Psychology Press.

Ertmer, P. A., & Newby, T. J. (1993). Behaviorism, cognitivism, constructivism: Comparing critical features from an instructional design perspective. *Performance improvement quarterly, 6*(4), 50-72.

Ewing, R. (2010). *The arts and Australian education: Realising potential*. Camberwell, Victoria: ACER Press.

Ferguson, D. (2001). Technology in a constructivist classroom. *Information Technology in Childhood Education Annual, 2001*(1), 45-55.

Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models : foundations, recent developments, and applications*. New York: Springer-Verlag.

Frisbie, D. A. (1988). Reliability of scores from teacher ‐ made tests. *Educational measurement: Issues and practice, 7*(1), 25-35.

Fullan, M., & Langworthy, M. (2013). *Towards a new end: New pedagogies for deep learning*. Seattle, WA: Creative Commons.

Gallaudet University. (n.d.). Instructions: Developing a Scoring Criteria (Rubrics). Retrieved from http://www.gallaudet.edu/office-of-academic-quality/assessment/assessment-of-student-learning/instructions-and-examples/developing-a-scoring-criteria-(rubrics)

Gershon, R. C. (2005). Computer adaptive testing. *Journal of applied measurement*.

Gill, T., & Bramley, T. (2008). *How accurate are examiners' judgments of script quality*.

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment*: Psychology Press.

Griffin, P., Care, E., & McGaw, B. (2012). The changing role of education and schools *Assessment and teaching of 21st century skills* (pp. 1-15): Springer.

Gunzenhauser, M. G. (2003). High-Stakes Testing and the Default Philosophy of Education. *Theory into Practice, 42*(1).

Haertel, E. H. (1999). Validity Arguments for High ‐ stakes Testing: In Search of the Evidence. *Educational measurement: Issues and practice, 18*(4), 5-9.

Harlen, W. (2005). Teachers' summative practices and assessment for learning–tensions and synergies. *Curriculum Journal, 16*(2), 207-223.

Hart, D. (1994). *Authentic assessment : a handbook for educators*. Menlo Park, Calif: Addison-Wesley Pub. Co.

Hattie, J. (2003). *Teachers Make a Difference, What is the research evidence?* Paper presented at the Australian Council for Educational Research: Annual Conference on Building Teacher Quality, Melbourne, Australia.

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

Henderson, M. (2014). Wicked problems and wicked designs in education and technology. Retrieved from http://newmediaresearch.educ.monash.edu.au/lnm/wicked-problems-and-wicked-designs-in-education-and-technology/

Henderson, M., Auld, G., Holkner, B., Russell, G., Seah, W. T., Fernando, A., & Romeo, G. (2010). Students creating digital video in the primary classroom : Student autonomy, learning outcomes, and professional learning communities. *Australian Council for Computers in Education, 24*(2), 9.

Henderson, M., Snyder, I., & Beale, D. (2013). Social media for collaborative learning: A review of school literature. *Australian Educational Computing, 28*(2).

Howe, K. R. a. M., M. S. . (1999). Ethics in educational research. *American Educational Research Association, 24*, 21-59.

Howell, J. (2013). *Teaching with ICT : Digital Pedagogies for Collaboration and Creativity*: Oxford University Press.

Howland, J. L., Jonassen, D. H., & Marra, R. M. (2012). *Meaningful learning with technology* (4th ed.). Boston: Pearson.

Hsu, Y.-C., Ching, Y.-H., & Grabowski, B. L. (2014). Web 2.0 applications and practices for learning through collaboration *Handbook of research on educational communications and technology* (pp. 747-758): Springer.

Humphry, S., & Heldsinger, S. (2009). *Do rubrics help to inform and direct teaching practices?* Paper presented at the From 2009-ACER Research Conference series.

Humphry, S. M., & Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher, 43*(5), 253-263.

Humphry, S. M., & McGrane, J. A. (2015). Equating a large-scale writing assessment using pairwise comparisons of performances. *The Australian Educational Researcher, 42*(4), 443-460.

Huygen, F. (1997). Report from Holland: Design criticism after postmodernism. *Design Issues, 13*(2), 40-43.

Hynes, P., & Younie, S. (2017). Bring your own device? *Debates in Computing and ICT Education*, 143.

International and Education Resource Network. (n.d.). iEARN. Retrieved from https://iearn.org/

Isaacs, T. (2013). *Key concepts in educational assessment*. London: SAGE.

Israel, M., & Hay, I. (2006). *Research ethics for social scientists : between ethical conduct and regulatory compliance*. London: Sage Publications.

ITU. (2012). *Overview of the internet of things*. Retrieved from

Jennings, J. L., & Bearak, J. M. (2014). “Teaching to the test” in the NCLB era: How test predictability affects our understanding of student performance. *Educational Researcher, 43*(8), 381-389.

JISC. (2010). *Effective Assessment in a Digital Age*. Bristol, UK.

Joint Information Systems Committee [JISC]. (2008). *Effective Practice with e-Portfolios: Supporting 21st century learning*. Bristol, UK: HEFCE.

Jonassen, D. (2014). Assessing problem solving. In J. M. Spector, D. M. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (pp. 269-288). New York, NY: Springer.

Jonassen, D. H. (2003). Using cognitive tools to represent problems. *Journal of Research on Technology in Education, 35*(3), 362-381.

Jonassen, D. H. (2006a). A constructivist's perspective on functional contextualism. *Educational Technology Research and Development, 54*(1), 43-47.

Jonassen, D. H. (2006b). *Modeling with technology : Mindtools for conceptual change* (3rd ed.). Upper Saddle River, N.J.: Pearson Merrill Prentice Hall.

Jones, I., & Alcock, L. (2012). Summative peer assessment of undergraduate calculus using adaptive comparative judgement. *Mapping university mathematics assessment practices*, 63-74.

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education, 39*(10), 1774-1787.

Jones, I., Swan, M., & Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education, 13*(1), 151-177.

Jones-Woodham, G. (2009). Using E-learning portfolio technology to support visual art learning. *Journal of Systemics, Cybernetics and Informatics, 7*.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*, 15.

Jordan, S. (2013). E-assessment: Past, present and future. *New Directions in the Teaching of Physical Sciences*(9), 87-106.

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research & Perspective, 2*(3), 135-170.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Westport, CT: American Council on Education.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342.

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice, 23*(2), 198-211.

Karagiorgi, Y., & Symeou, L. (2005). Translating constructivism into instructional design: Potential and limitations. *Journal of Educational Technology & Society, 8*(1).

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard ‐ setting method: A literature review. *Educational measurement: Issues and practice, 25*(1), 4-12.

Kaufman, T. E., Graham, C. R., Picciano, A. G., Popham, J. A., & Wiley, D. (2014). Data-driven decision making in the K-12 classroom. In J. M. Spector, D. M. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (pp. 337-346). New York, NY: Springer.

Keane, T., & Keane, W. (2017). Achievements and challenges: Implementing a 1: 1 program in a secondary school. *Education and Information Technologies, 22*(3), 1025-1041.

Kimbell, R. (2007). *Technology and the assessment of creative performance*. Paper presented at the Cambridge Assessment Conference, London.

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

Kimbell, R. (2011). Handle with care. *Design and Technology Education: An International Journal, 16*(1).

Kimbell, R., & Stables, K. (2007). *Researching design learning : issues and findings from two decades of research and development*. Dordrecht, The Netherlands: Springer.

Kimbell, R., Wheeler, T., Stables, K., Sheppard, T., Martin, D. D., Pollitt, A., & Whitehouse, G. (2009). *E-scape portfolio assessment: phase 3 report.* Retrieved from London, UK:

Knight, P. (2001). *A briefing on key concepts: Formative and summative, criterion and norm-referenced assessment*

Koehler, M. J. (2017). TPACK Explained. Retrieved from http://matt-koehler.com/tpack2/tpack-explained/

Koehler, M. J., Mishra, P., & Cain, W. (2013). What is technological pedagogical content knowledge (TPACK)? *Journal of Education*, 13-19.

Kohn, A. (2000). *The case against standardized testing: Raising the scores, ruining the schools*: Heinemann Portsmouth, NH.

Kohn, A. (2006). The trouble with rubrics. *English journal, 95*(4), 12-15.

Koretz, D. (1998). Large ‐ scale portfolio assessments in the US: evidence pertaining to the quality of measurement. *Assessment in Education, 5*(3), 309-334.

Kubiszyn, T., & Borich, G. D. (1993). *Educational testing and measurement : classroom application and practice* (4th ed.). New York: HarperCollins College Publishers.

Laming, D. (2011). *Human judgment: the eye of the beholder*. Hampshire: Cengage Learning.

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema, 26*(1).

Lim, C. P., Zhao, Y., Tondeur, J., Chai, C. S., & Chin-Chung, T. (2013). Bridging the gap: Technology trends and use of technology in schools. *Journal of Educational Technology & Society, 16*(2).

Lin, H., & Dwyer, F. (2006). The fingertip effects of computer-based assessment in education. *TechTrends, 50*(6), 27-31.

Lindenberg, K., Schoenmaekers, S., Halasy, K., & Rehbein, F. (2017). OP-67: School-related risk factors associated with Internet gaming disorder and internet addiction. *Journal of Behavioral Addictions, 6*(S1), 32-33.

Lockee, B. B., & Wang, F. (2014). Visual arts education. In J. M. Spector, D. M. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (pp. 583-590). New York, NY: Springer.

Lund, J. (1997). Authentic assessment: Its development & applications. *Journal of Physical Education, Recreation & Dance, 68*(7), 25-28.

Madaus, G. F., & O'Dwyer, L. M. (1999). A short history of performance assessment: Lessons learned. *Phi Delta Kappan, 80*(9), 688.

Madeja, S. S., Dorn, C. M., & Sabol, F. R. (2004). Alternative assessment strategies for schools. *Arts education policy review, 105*(5), 3.

Marchant, G. J., & Finch, W. H. (2016). Student, school, and country: The relationship of SES and inequality to achievement. *Journal of Global Research in Education and Social Science, 6*(4), 187-196.

Mareschal, D., Butterworth, B., & Tolmie, A. (2014). *Educational neuroscience*. Chichester: Wiley Blackwell.

Martindale, T., & Dowdy, M. (2016). Issues in research, design, and development of personal learning environments. In T. Anderson (Ed.), *Emergence and innovation in digiital learning* (pp. 119-142). Edmonton: Athabasca University.

Masek, M., Murcia, K., Morrison, J., Newhouse, C. P., & Hackling, M. (2012). *Learning in transformational computer games: Exploring design principles for a nanotechnology game.* Paper presented at the Australian Association for Research in Education, Sydney.

Masters, G. N. (2013). *Reforming educational assessment: Imperatives, principles and challenges*. Retrieved from Camberwell, Victoria:

McLoughlin, C., & Lee, M. J. (2007). *Social software and participatory learning: Pedagogical choices with technology affordances in the Web 2.0 era.* Paper presented at the ICT: Providing choices for learners and learning. Proceedings ascilite Singapore 2007.

McMahon, S., & Jones, I. (2015). A comparative judgement approach to teacher assessment. *Assessment in Education: Principles, Policy & Practice, 22*(3), 368-389.

Merrill, D. M. (1992). Constructivism and Instructional Design. In T. M. Duffy & D. Jonassen (Eds.), *Constructivism and the Technology of Instruction*. Hillsdale, N.J. ; London: Lawrence Erlbaum.

Messick, S. (1980). Test validity and the ethics of assessment. *The American Psychologist, 35*(11), 16.

Messick, S. (1987). Validity. *ETS Research Report Series, 1987*(2), i-208.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 7.

Messick, S. (1990). Validity of test interpretation and use. *ETS Research Report Series, 1990*(1), 1487-1495.

Messick, S. (1993a). Foundations of validity: Meaning and consequences in psychological assessment. *ETS Research Report Series, 1993*(2), i-18.

Messick, S. (1993b). Validity. In R. L. Linn (Ed.), *Educational Measurement, 3rd Ed.* New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 10.

Messick, S. (1996). Validity and washback in language testing. *ETS Research Report Series, 1996*(1).

Miller, D. G. (2011). *An investigation into the feasibility of using digital representations of students' work for authentic and reliable performance assessment in Applied*

Information Technology. (Doctoral dissertation), Edith Cowan University, Perth, Western Australia.

Miller, D. M., & Linn, R. L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement, 24*(4), 367-378.

Miller, M. D., Linn, R. L., Gronlund, N. E., & Linn, R. L. (2009). *Measurement and assessment in teaching* (10th ed.). Upper Saddle River, N.J.: Merrill/Pearson.

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers college record, 108*(6), 1017.

Moore, C. (2016). The Future of Work: What Google Shows Us About the Present and Future of Online Collaboration. *TechTrends, 60*(3), 233-244.

Moskal, B. M. (2000). Scoring Rubrics: What, When and How? *Practical Assessment, Research & Evaluation, 7*(3). Retrieved from https://scholar.google.com.au/scholar?q=moskal+scoring+rubrics%3A+what&btnG=&hl=en&as_sdt=0%2C5

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation, 7*(10). Retrieved from http://pareonline.net/getvn.asp?v=7&n=10

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education*, 109-162.

NAPLAN. (2013). NAPLAN Online. Retrieved from http://www.nap.edu.au/online-assessment/naplan-online/naplan-online.html

Newhouse, C. P. (2011a). [Proposal]. Authentic digital representation of creative works in education: Addressing the challenges of digitisation and assessment. Perth, Western Australia: Australian Research Council.

Newhouse, C. P. (2011b). Using IT to assess IT: Towards greater authenticity in summative performance assessment. *Computers & Education, 56*(2), 388-402.

Newhouse, C. P. (2014). Using digital representations of practical production work for summative assessment. *Assessment in Education: Principles, Policy & Practice, 21*(2), 205-220. doi:10.1080/0969594x.2013.868341

Newhouse, C. P. (2017). STEM the Boredom: Engage Students in the Australian Curriculum Using ICT with Problem-Based Learning and Assessment. *Journal of Science Education and Technology, 26*(1), 44-57.

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*. Retrieved from Perth, Western Australia:

Newhouse, C. P., & Tarricone, P. (2014). Digitizing Practical Production Work for High-Stakes Assessments. *Canadian Journal of Learning and Technology, 40*(2), n2.

Newhouse, C. P., Trinidad, S., & Clarkson, B. (2002). *Quality pedagogy and effective learning with Information and Communications Technologies (ICT): A review of the literature*. Retrieved from Perth, Western Australia:

Newhouse, C. P., Williams, J., Penney, D., Pagram, J., Jones, A., Campbell, A., & Cooper, M. (2011). *Digital Forms of Assessment Final Report*. Retrieved from Perth, Australia:

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools*: Harvard Education Press Cambridge, MA.

No More Marking. (2017). Retrieved from https://www.nomoremarking.com/

Orfield, G., & Kornhaber, M. L. (2001). *Raising standards or raising barriers?: Inequality and high-stakes testing in public education*: Century Foundation Press New York.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks: Sage Publications.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know : the science and design of educational assessment*. Washington (D.C.): National academy Press.

Perkins, D. N. (1992a). Constructivism and Instructional Design. In T. M. Duffy & D. Jonassen (Eds.), *Constructivism and the Technology of Instruction*. Hillsdale, N.J. ; London: Lawrence Erlbaum.

Perkins, D. N. (1992b). Technology meets constructivism: Do they make a marriage? In T. M. Duffy & D. Jonassen (Eds.), *Constructivism and the technology of instruction: A conversation*. New Jersey: Lawrence Erlbaum Associates.

Perlman, C. (2003). *Performance assessment: Designing appropriate performance tasks and scoring rubrics* J. E. Wall & G. R. Walz (Eds.), *Measuring Up: Assessment Issues for Teachers, Counselors, adn Administrators* Retrieved from https://eric.ed.gov/?q=Measuring+Up%3a+Assessment+Issues+for+Teachers%2c+Counselors%2c+and+Administrators.&id=ED480379

Pollitt, A. (2004). *Let's stop marking exams.* Paper presented at the IAEA Conference, Philadelphia.

Pollitt, A. (2012a). Comparative judgement for assessment. *International Journal of Technology and Design Education, 22*(2), 157-170.

Pollitt, A. (2012b). The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281-300. doi:10.1080/0969594x.2012.665354

Pollitt, A. (2012c). Validity Cannot Be Created, It Can Only Be Lost. *Measurement: Interdisciplinary Research & Perspective, 10*(1-2), 100-103. doi:10.1080/15366367.2012.686868

Pollitt, A., & Whitehouse, C. (2012). Using adaptive comparative judgement to obtain a highly reliable rank order in summative assessment.

Popham, W. J. (1997). What's wrong-and what's right-with rubrics. *Educational leadership, 55*, 72-75.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York: Routledge.

Rayment, T. (1999). Assessing National Curriculum Art AT2, Knowledge and Understanding: A small ‐ scale project at Key Stage 3. *International Journal of Art & Design Education, 18*(2), 189-194.

Rayment, T. (2007). Introduction: The problem of assessment in art and design. In T. Rayment (Ed.), *The problem of assessment in art and design* (pp. 7-9). Bristol, UK: Intellect Books.

Rayment, T., & Britton, B. (2007). The assessment of GCSE art: Criterion-referencing and cognitive abilities. In T. Rayment (Ed.), *The problem of assessment in art and design* (pp. 41-48). Bristol, UK: Intellect Books.

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education, 35*(4), 435-448.

Redecker, C., & Johannessen, Ø. (2013). Changing assessment—Towards a new assessment paradigm using ICT. *European Journal of Education, 48*(1), 79-96.

Ridgway, J., McCusker, S., & Pead, D. (2004). *Literature review of e-assessment*: Futurelab.

Roads, C. (2015). *Composing electronic music: a new aesthetic*: Oxford University Press, USA.

Rust, J., & Golombok, S. (2014). *Modern psychometrics: The science of psychological assessment*. New York, NY: Routledge.

Sadler, D. R. (1986). *Subjectivity, Objectivity, and Teachers' Qualitative Judgements (Discussion Paper 5)*. Brisbane, Australia: Board of Senior Secondary School Studies. Queensland Studies Authority.

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education, 13*(2), 191-209.

Sadler, D. R. (2009). Transforming holistic assessment and grading into a vehicle for complex learning *Assessment, learning and judgement in higher education* (pp. 1-19): Springer.

Scardamalia, M., Bransford, J., Kozma, B., & Quellmalz, E. (2012). New assessments and environments for knowledge building *Assessment and teaching of 21st century skills* (pp. 231-300): Springer.

School Curriculum and Standards Authority. (2014a). OLNA. Retrieved from http://wace1516.scsa.wa.edu.au/assessment/olna

School Curriculum and Standards Authority. (2014b). Technologies. Retrieved from https://senior-secondary.scsa.wa.edu.au/syllabus-and-support-materials/technologies

Scriven, M. S. (1967). The methodology of evaluation (Perspectives of Curriculum Evaluation, and AERA monograph Series on Curriculum Evaluation, No. 1). *Chicago: Rand NcNally*.

Selinger, M., & Kaye, L. (2005). ICT tools and applications. *Learning to Teach Using ICT in the Secondary School: A Companion to School Experience*, 74.

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.

Shrader-Frechette, K. S. (1994). *Ethics of scientific research*: Rowman & Littlefield.

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher, 36*(8), 477-481.

Sireci, S. G. (Ed.) (2009). *Packing and unpacking sources of validity evidence: History repeats itself again*: IAP.

Skryabin, M., Zhang, J., Liu, L., & Zhang, D. (2015). How the ICT development level and usage influence student achievement in reading, mathematics, and science. *Computers & Education, 85*, 49-58.

Spector, J. M. (2014). Conceptualizing the emerging field of smart learning environments. *Smart learning environments, 1*(1), 2.

Stables, K. (2015). Assessment: Feedback from Our Pasts, Feedforward for Our Futures.

Stables, K. (2017a). Critiquing Design: Perspectives and World Views on Design and Design and Technology Education, for the Common Good. In P. J. Williams & K. Stables (Eds.), *Critique in Design and Technology Education* (pp. 51-70). Singapore: Springer Singapore.

Stables, K. (2017b). *Holistic approaches to learning, teaching and assessment in Technology Education: Authenticity, challenges, supportive approaches and tools.* Paper presented at the USBTENZ-ICTE Conference, Christchurch, New Zealand.

Stacey, K., & Wiliam, D. (2012). Technology and assessment in mathematics *Third international handbook of mathematics education* (pp. 721-751): Springer.

Steinkuehler, C., Squire, K., & Barab, S. (2012). *Games, learning, and society: Learning and meaning in the digital age*: Cambridge University Press.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4). Retrieved from http://www.pareonline.net/getvn.asp?v=9&n=4

Stobart, G. (2010). *Assessment fit-for-the-future.* Paper presented at the International Association for Educational Assessment, Bangkok, Thailand.

Stowell, L., & McDaniel, J. (1997). The changing face of assessment. *What current research says to the middle level practitioner*, 137-150.

Taras, M. (2005). Assessment–summative and formative–some theoretical reflections. *British Journal of Educational Studies, 53*(4), 466-478.

Tatum, D. S. (2000). Rasch analysis: an introduction to objective measurement. *Laboratory Medicine, 31*(5), 272-274.

Taylor, P. (2006). *Assessment in arts education*. Portsmouth, NH: Heinemann.

The Cross Sectoral Assessment Working Party. (2011). *Teachers' Guide to Assessment*. ACT Retrieved from http://www.det.act.gov.au/__data/assets/pdf_file/0011/297182/Teachers_Guide_to_Assessment_Web.pdf.

Thorndike, R. L., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education*. Boston, MA: Pearson.

Thorndike, R. M. (1997). *Measurement and evaluation in psychology and education* (6th ed.). Upper Saddle River, N.J: Merrill.

Thurstone, L. L. (1927). A Law of Comparative Judgement. *Psychological Review, 34*, 14.

Thurstone, L. L. (1928). Attitudes can be measured. *American journal of sociology, 33*(4), 529-554.

Timmis, S., Broadfoot, P., Sutherland, R., & Oldfield, A. (2016). Rethinking assessment in a digital age: Opportunities, challenges and risks. *British Educational Research Journal, 42*(3), 454-476.

Traub, R. E., & Rowley, G. L. (1991) Understanding Reliability. *Instructional Topics in Educational Measurement.*

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*: Harvard university press.

Ward, A. W., Stoker, H. W., & Murray-Ward, M. (1996). *Educational measurement, origins, theories and explications*. Lanham, Md: University Press of America.

Wenger, E. (2011). Communities of practice: A brief introduction.   Retrieved from https://scholarsbank.uoregon.edu/xmlui/bitstream/handle/1794/11736/A brief introduction to CoP.pdf?sequence%E2%80%B0=%E2%80%B01

Whitehouse, C., & Pollitt, A. (2012). Using adaptive comparative judgement to obtain a highly reliable rank order in summative assessment. *Manchester: AQA Centre for Education Research and Policy*.

Wiggins, G. (1990). The Case for Authentic Assessment. ERIC Digest.

Wiggins, G. (2011). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 92*(7), 81-93.

Wiggins, J. (2015). Constructivism, policy, and arts education. *Arts education policy review, 116*(3), 3.

Wiliam, D. (1994). *Towards a philosophy for educational assessment.* Paper presented at the an update on a paper given at the British Educational Research Association's 20th Annual conference in Oxford in.

Wiliam, D. (2000). *Integrating formative and summative functions of assessment.* Paper presented at the Working group.

Williams, P. J. (2000). Design: The only methodology of technology?

Williams, P. J. (2009). Assessment of student performance in engineering.

Wilson, M., Scalise, K., & Gochyyev, P. (2015). Rethinking ICT literacy: From computer skills to social network settings. *Thinking Skills and Creativity, 18*, 65-80.

Wilson, S. (2017). Reaching Full Digitization in the Classroom. *The Education Digest, 83*(3), 61.

Wolf, D. P. (1989). Portfolio assessment: Sampling student work. *Educational leadership, 46*(7), 35-35.

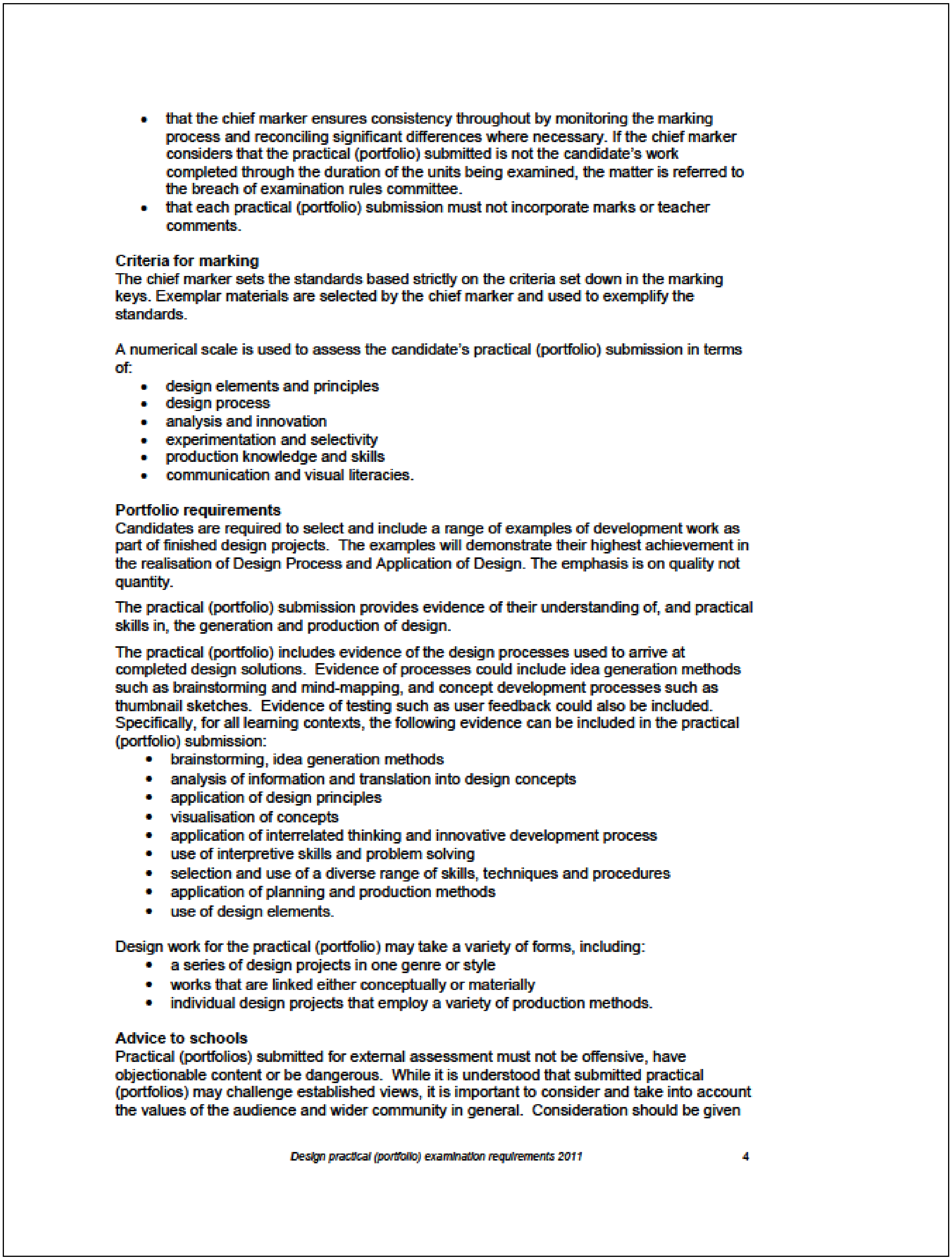
Wooff, D., Bell, D., & Owen-Jackson, G. (2013). Assessment questions. *Owen-Jackson, G.'Debates in design and technology'Abingdon: Routledge*, 180-192.

Wyatt‐Smith, C., Klenowski, V., & Gunn, S. (2010). The centrality of teachers' judgement practice in assessment: A study of standards in moderation. *Assessment in Education: Principles, Policy & Practice, 17*(1), 59-75.

Yu, C. H. (2011). A simple guide to the item response theory (IRT) and Rasch modeling*14*. Retrieved from http://www.creative-wisdom.com/computer/sas/IRT_tutorial.doc

APPENDICES

Appendix A Portfolio Requirements and Design Brief (Design)



- that the chief marker ensures consistency throughout by monitoring the marking process and reconciling significant differences where necessary. If the chief marker considers that the practical (portfolio) submitted is not the candidate's work completed through the duration of the units being examined, the matter is referred to the breach of examination rules committee.
- that each practical (portfolio) submission must not incorporate marks or teacher comments.

Criteria for marking
The chief marker sets the standards based strictly on the criteria set down in the marking keys. Exemplar materials are selected by the chief marker and used to exemplify the standards.

A numerical scale is used to assess the candidate's practical (portfolio) submission in terms of:

- design elements and principles
- design process
- analysis and innovation
- experimentation and selectivity
- production knowledge and skills
- communication and visual literacies.

Portfolio requirements
Candidates are required to select and include a range of examples of development work as part of finished design projects. The examples will demonstrate their highest achievement in the realisation of Design Process and Application of Design. The emphasis is on quality not quantity.

The practical (portfolio) submission provides evidence of their understanding of, and practical skills in, the generation and production of design.

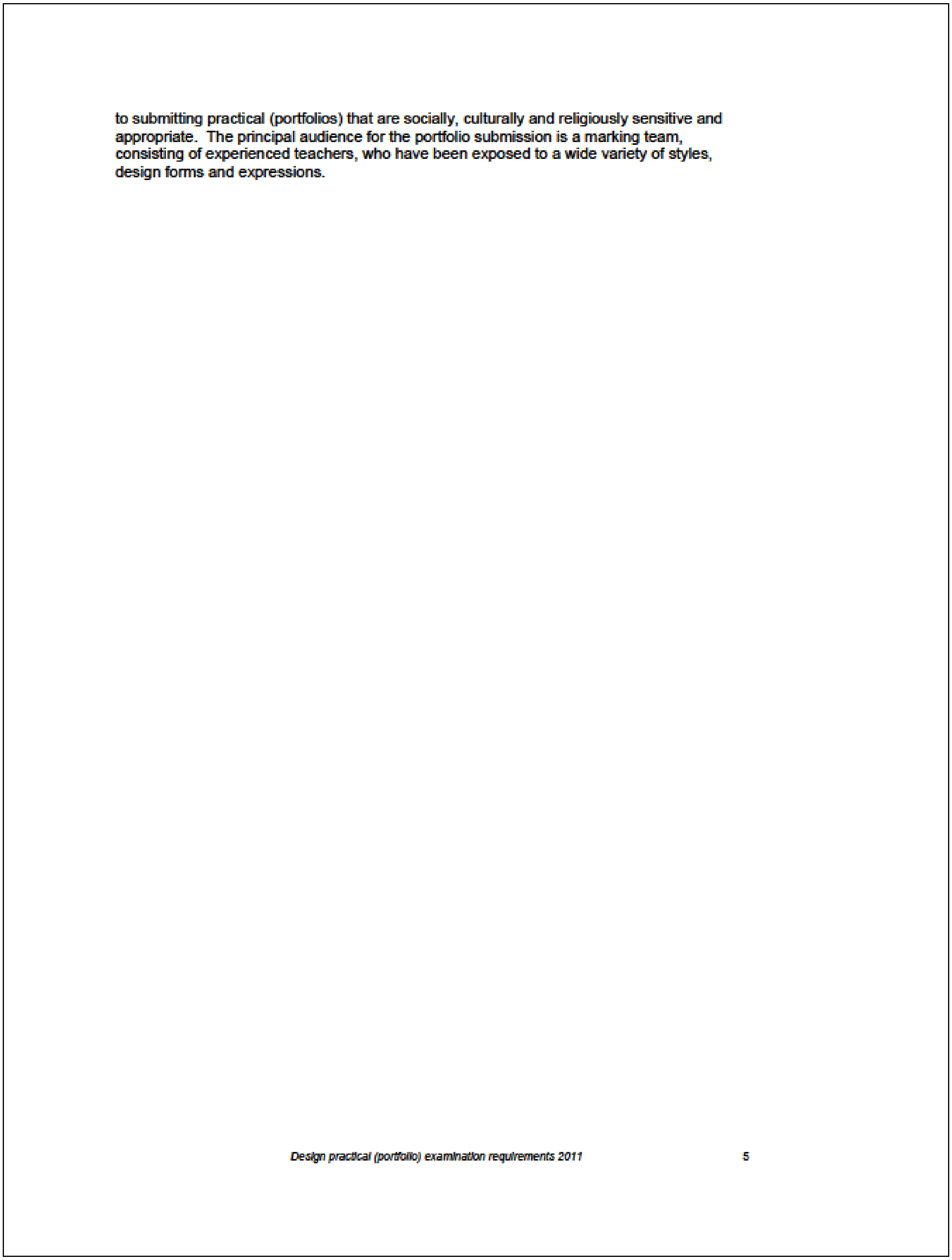
The practical (portfolio) includes evidence of the design processes used to arrive at completed design solutions. Evidence of processes could include idea generation methods such as brainstorming and mind-mapping, and concept development processes such as thumbnail sketches. Evidence of testing such as user feedback could also be included. Specifically, for all learning contexts, the following evidence can be included in the practical (portfolio) submission:

- brainstorming, idea generation methods
- analysis of information and translation into design concepts
- application of design principles
- visualisation of concepts
- application of interrelated thinking and innovative development process
- use of interpretive skills and problem solving
- selection and use of a diverse range of skills, techniques and procedures
- application of planning and production methods
- use of design elements.

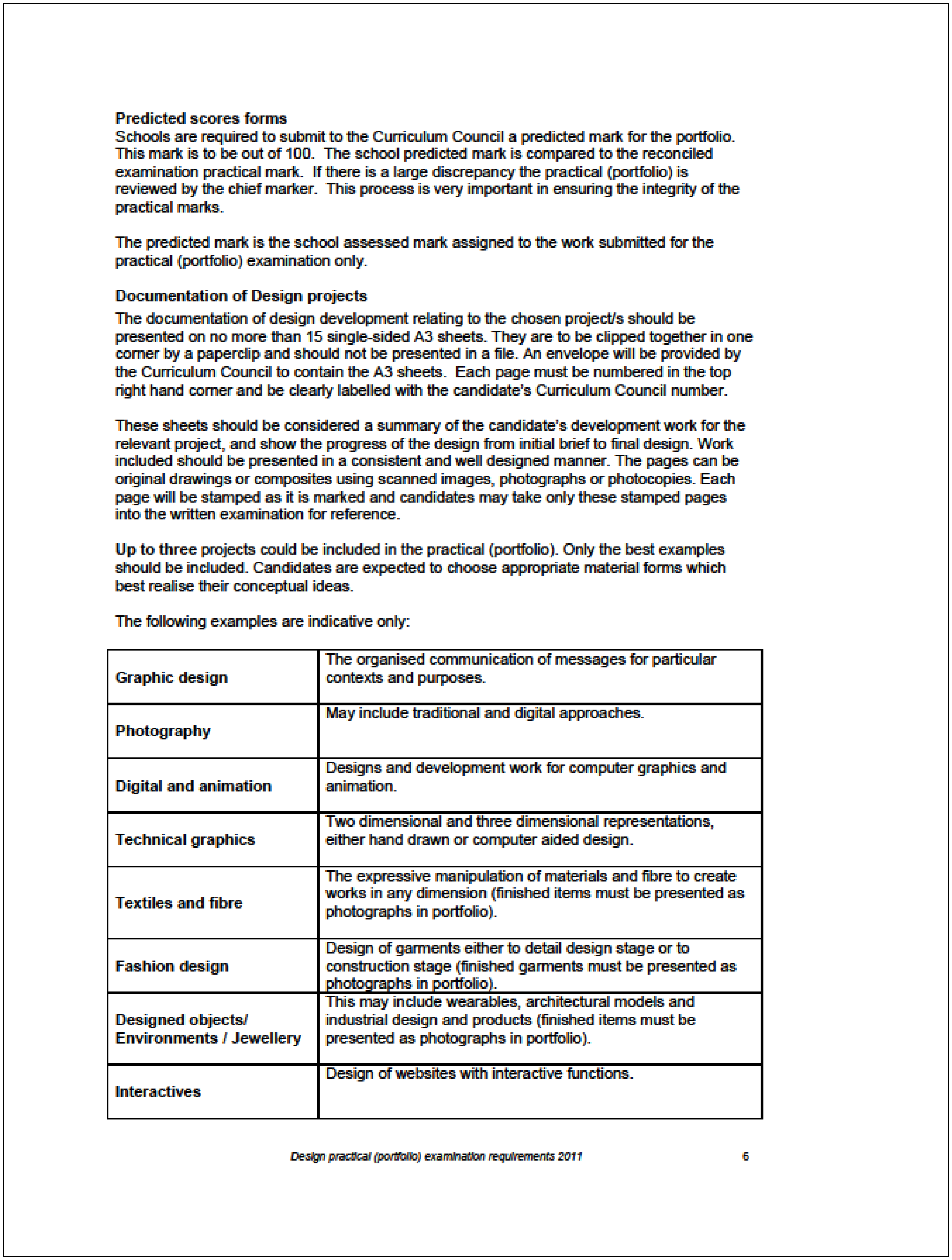
Design work for the practical (portfolio) may take a variety of forms, including:

- a series of design projects in one genre or style
- works that are linked either conceptually or materially
- individual design projects that employ a variety of production methods.

Advice to schools
Practical (portfolios) submitted for external assessment must not be offensive, have objectionable content or be dangerous. While it is understood that submitted practical (portfolios) may challenge established views, it is important to consider and take into account the values of the audience and wider community in general. Consideration should be given



to submitting practical (portfolios) that are socially, culturally and religiously sensitive and appropriate. The principal audience for the portfolio submission is a marking team, consisting of experienced teachers, who have been exposed to a wide variety of styles, design forms and expressions.



Predicted scores forms

Schools are required to submit to the Curriculum Council a predicted mark for the portfolio. This mark is to be out of 100. The school predicted mark is compared to the reconciled examination practical mark. If there is a large discrepancy the practical (portfolio) is reviewed by the chief marker. This process is very important in ensuring the integrity of the practical marks.

The predicted mark is the school assessed mark assigned to the work submitted for the practical (portfolio) examination only.

Documentation of Design projects

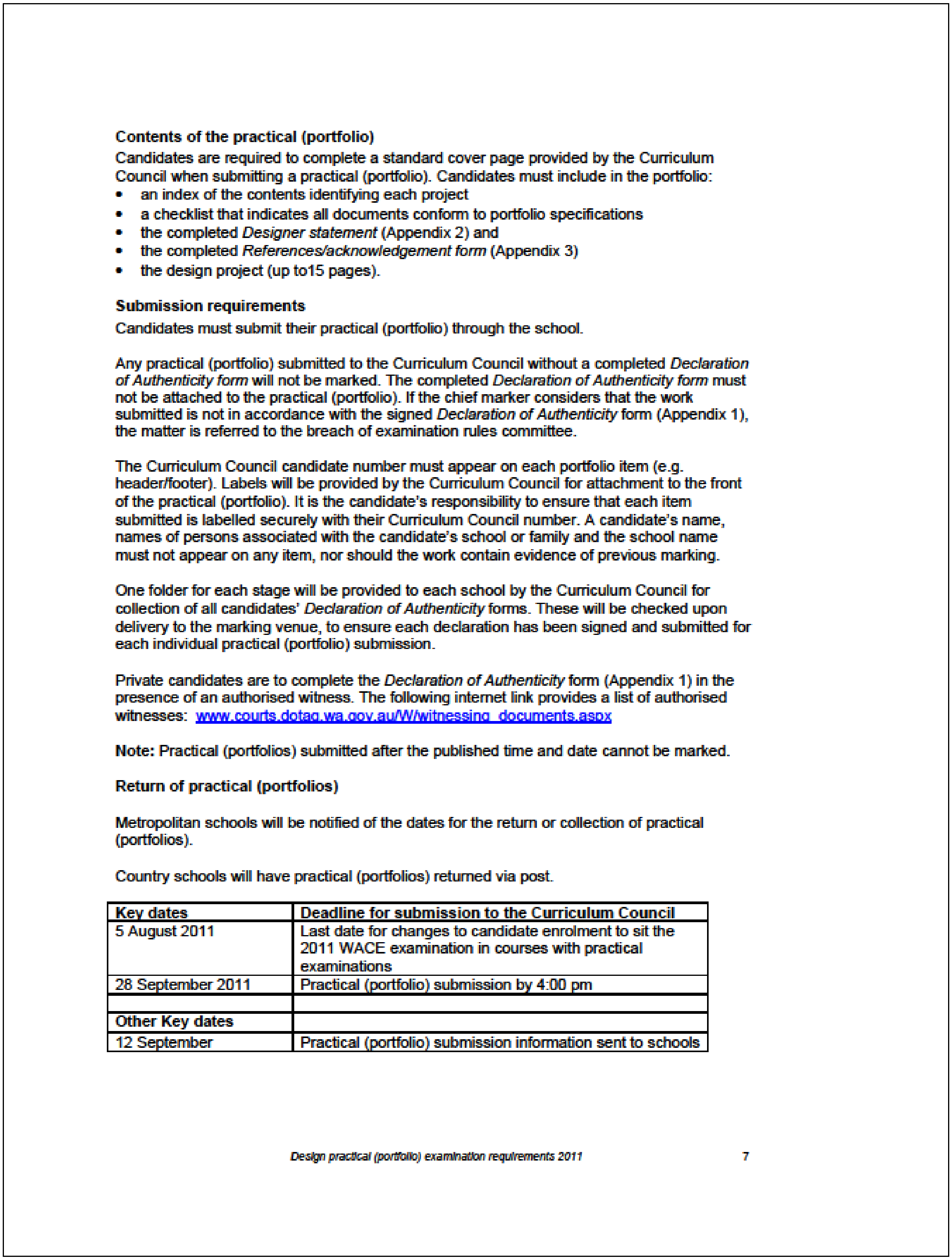
The documentation of design development relating to the chosen project/s should be presented on no more than 15 single-sided A3 sheets. They are to be clipped together in one corner by a paperclip and should not be presented in a file. An envelope will be provided by the Curriculum Council to contain the A3 sheets. Each page must be numbered in the top right hand corner and be clearly labelled with the candidate's Curriculum Council number.

These sheets should be considered a summary of the candidate's development work for the relevant project, and show the progress of the design from initial brief to final design. Work included should be presented in a consistent and well designed manner. The pages can be original drawings or composites using scanned images, photographs or photocopies. Each page will be stamped as it is marked and candidates may take only these stamped pages into the written examination for reference.

Up to three projects could be included in the practical (portfolio). Only the best examples should be included. Candidates are expected to choose appropriate material forms which best realise their conceptual ideas.

The following examples are indicative only:

Graphic design	The organised communication of messages for particular contexts and purposes.
Photography	May include traditional and digital approaches.
Digital and animation	Designs and development work for computer graphics and animation.
Technical graphics	Two dimensional and three dimensional representations, either hand drawn or computer aided design.
Textiles and fibre	The expressive manipulation of materials and fibre to create works in any dimension (finished items must be presented as photographs in portfolio).
Fashion design	Design of garments either to detail design stage or to construction stage (finished garments must be presented as photographs in portfolio).
Designed objects/ Environments / Jewellery	This may include wearables, architectural models and industrial design and products (finished items must be presented as photographs in portfolio).
Interactives	Design of websites with interactive functions.



Contents of the practical (portfolio)

Candidates are required to complete a standard cover page provided by the Curriculum Council when submitting a practical (portfolio). Candidates must include in the portfolio:

- an index of the contents identifying each project
- a checklist that indicates all documents conform to portfolio specifications
- the completed *Designer statement* (Appendix 2) and
- the completed *References/acknowledgement form* (Appendix 3)
- the design project (up to15 pages).

Submission requirements

Candidates must submit their practical (portfolio) through the school.

Any practical (portfolio) submitted to the Curriculum Council without a completed *Declaration of Authenticity form* will not be marked. The completed *Declaration of Authenticity form* must not be attached to the practical (portfolio). If the chief marker considers that the work submitted is not in accordance with the signed *Declaration of Authenticity* form (Appendix 1), the matter is referred to the breach of examination rules committee.

The Curriculum Council candidate number must appear on each portfolio item (e.g. header/footer). Labels will be provided by the Curriculum Council for attachment to the front of the practical (portfolio). It is the candidate's responsibility to ensure that each item submitted is labelled securely with their Curriculum Council number. A candidate's name, names of persons associated with the candidate's school or family and the school name must not appear on any item, nor should the work contain evidence of previous marking.

One folder for each stage will be provided to each school by the Curriculum Council for collection of all candidates' *Declaration of Authenticity* forms. These will be checked upon delivery to the marking venue, to ensure each declaration has been signed and submitted for each individual practical (portfolio) submission.

Private candidates are to complete the *Declaration of Authenticity* form (Appendix 1) in the presence of an authorised witness. The following internet link provides a list of authorised witnesses: www.courts.dotag.wa.gov.au/W/witnessing_documents.aspx

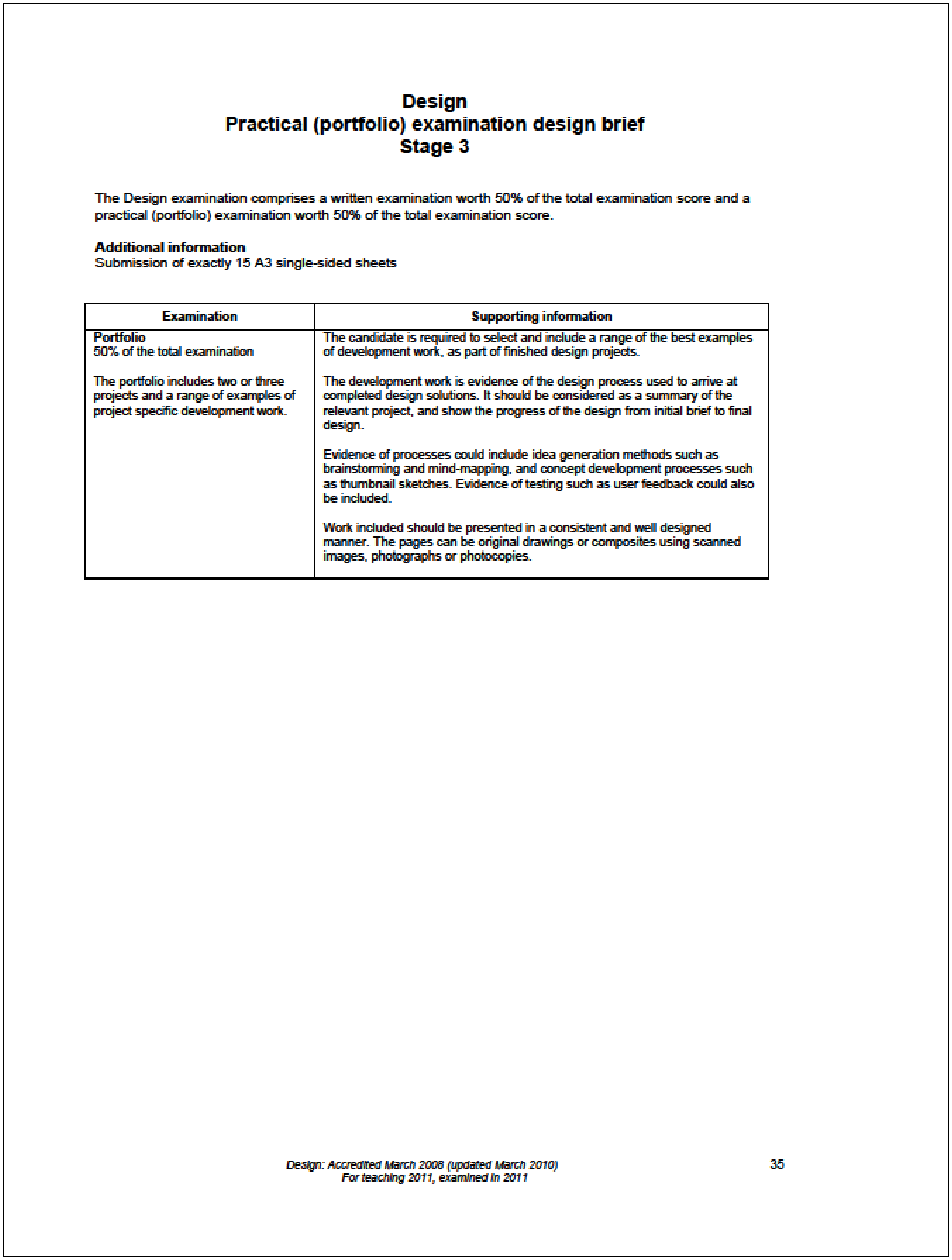
Note: Practical (portfolios) submitted after the published time and date cannot be marked.

Return of practical (portfolios)

Metropolitan schools will be notified of the dates for the return or collection of practical (portfolios).

Country schools will have practical (portfolios) returned via post.

Key dates	**Deadline for submission to the Curriculum Council**
5 August 2011	Last date for changes to candidate enrolment to sit the 2011 WACE examination in courses with practical examinations
28 September 2011	Practical (portfolio) submission by 4:00 pm
Other Key dates	
12 September	Practical (portfolio) submission information sent to schools



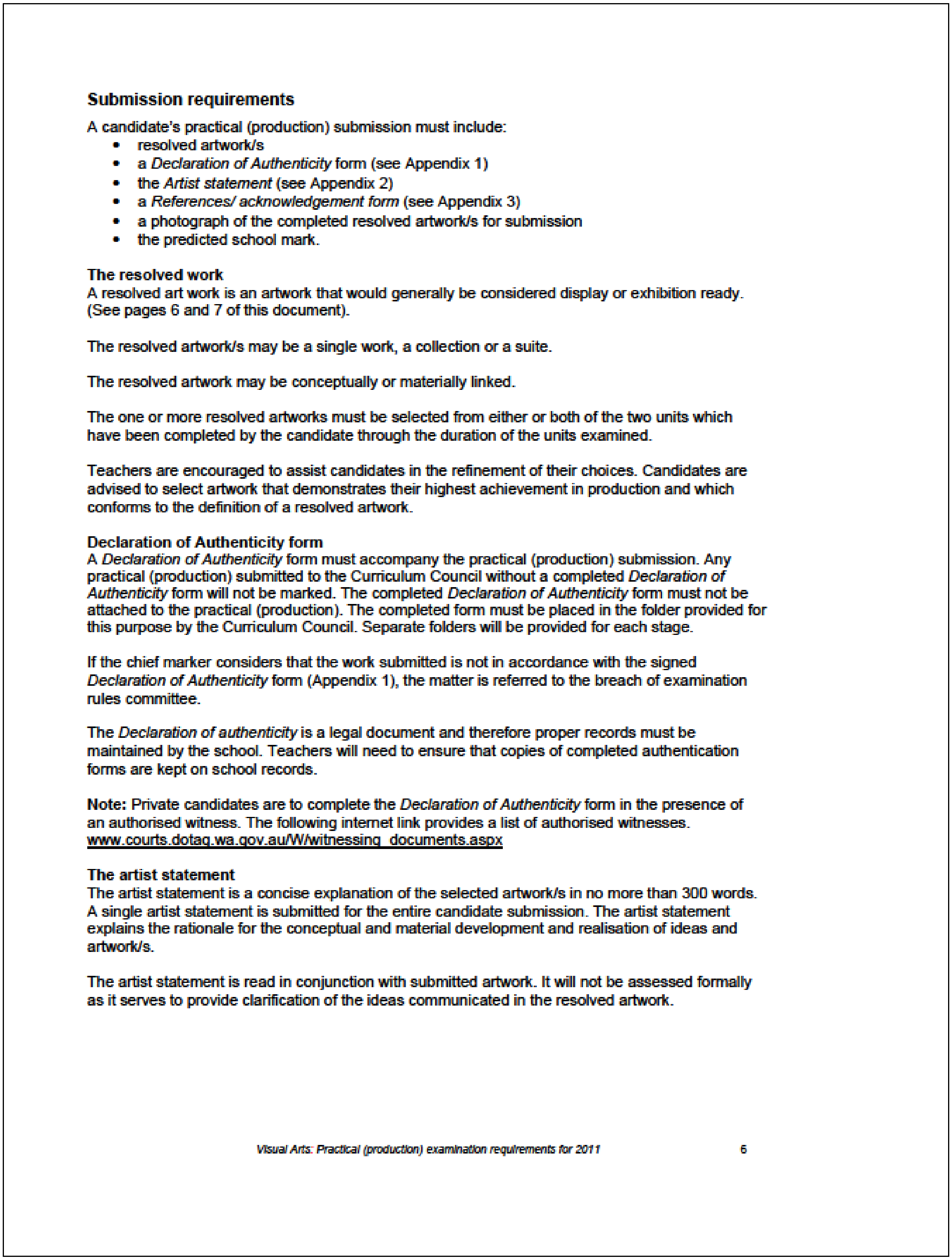
Design
Practical (portfolio) examination design brief
Stage 3

The Design examination comprises a written examination worth 50% of the total examination score and a practical (portfolio) examination worth 50% of the total examination score.

Additional information
Submission of exactly 15 A3 single-sided sheets

Examination	Supporting information
Portfolio 50% of the total examination The portfolio includes two or three projects and a range of examples of project specific development work.	The candidate is required to select and include a range of the best examples of development work, as part of finished design projects. The development work is evidence of the design process used to arrive at completed design solutions. It should be considered as a summary of the relevant project, and show the progress of the design from initial brief to final design. Evidence of processes could include idea generation methods such as brainstorming and mind-mapping, and concept development processes such as thumbnail sketches. Evidence of testing such as user feedback could also be included. Work included should be presented in a consistent and well designed manner. The pages can be original drawings or composites using scanned images, photographs or photocopies.

Appendix B Portfolio Requirements and Design Brief (Visual Arts)



Submission requirements

A candidate's practical (production) submission must include:

- resolved artwork/s
- a *Declaration of Authenticity* form (see Appendix 1)
- the *Artist statement* (see Appendix 2)
- a *References/ acknowledgement form* (see Appendix 3)
- a photograph of the completed resolved artwork/s for submission
- the predicted school mark.

The resolved work

A resolved art work is an artwork that would generally be considered display or exhibition ready. (See pages 6 and 7 of this document).

The resolved artwork/s may be a single work, a collection or a suite.

The resolved artwork may be conceptually or materially linked.

The one or more resolved artworks must be selected from either or both of the two units which have been completed by the candidate through the duration of the units examined.

Teachers are encouraged to assist candidates in the refinement of their choices. Candidates are advised to select artwork that demonstrates their highest achievement in production and which conforms to the definition of a resolved artwork.

Declaration of Authenticity form

A *Declaration of Authenticity* form must accompany the practical (production) submission. Any practical (production) submitted to the Curriculum Council without a completed *Declaration of Authenticity* form will not be marked. The completed *Declaration of Authenticity* form must not be attached to the practical (production). The completed form must be placed in the folder provided for this purpose by the Curriculum Council. Separate folders will be provided for each stage.

If the chief marker considers that the work submitted is not in accordance with the signed *Declaration of Authenticity* form (Appendix 1), the matter is referred to the breach of examination rules committee.

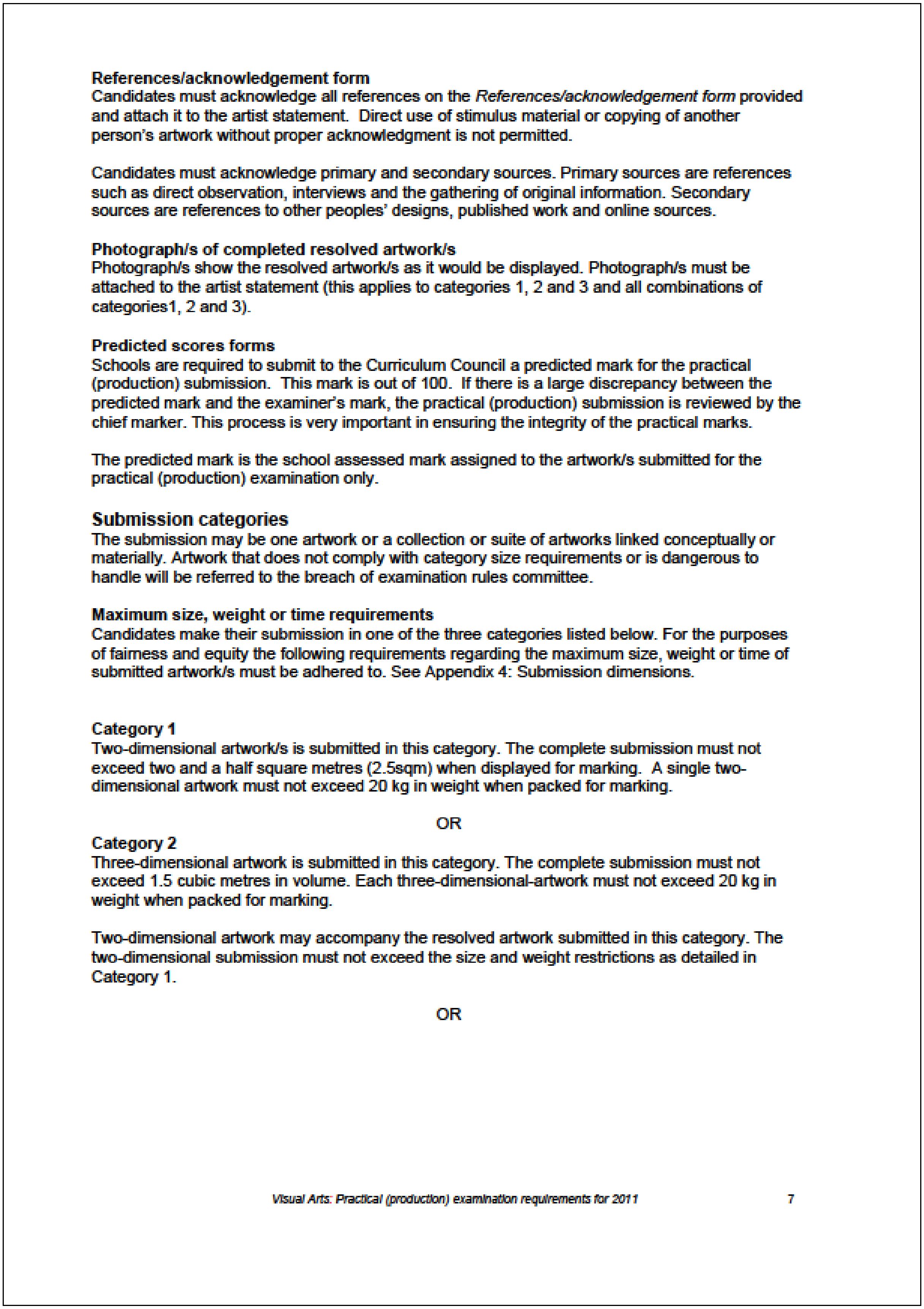
The *Declaration of authenticity* is a legal document and therefore proper records must be maintained by the school. Teachers will need to ensure that copies of completed authentication forms are kept on school records.

Note: Private candidates are to complete the *Declaration of Authenticity* form in the presence of an authorised witness. The following internet link provides a list of authorised witnesses. www.courts.dotag.wa.gov.au/W/witnessing_documents.aspx

The artist statement

The artist statement is a concise explanation of the selected artwork/s in no more than 300 words. A single artist statement is submitted for the entire candidate submission. The artist statement explains the rationale for the conceptual and material development and realisation of ideas and artwork/s.

The artist statement is read in conjunction with submitted artwork. It will not be assessed formally as it serves to provide clarification of the ideas communicated in the resolved artwork.



References/acknowledgement form
Candidates must acknowledge all references on the *References/acknowledgement form* provided and attach it to the artist statement. Direct use of stimulus material or copying of another person's artwork without proper acknowledgment is not permitted.

Candidates must acknowledge primary and secondary sources. Primary sources are references such as direct observation, interviews and the gathering of original information. Secondary sources are references to other peoples' designs, published work and online sources.

Photograph/s of completed resolved artwork/s
Photograph/s show the resolved artwork/s as it would be displayed. Photograph/s must be attached to the artist statement (this applies to categories 1, 2 and 3 and all combinations of categories1, 2 and 3).

Predicted scores forms
Schools are required to submit to the Curriculum Council a predicted mark for the practical (production) submission. This mark is out of 100. If there is a large discrepancy between the predicted mark and the examiner's mark, the practical (production) submission is reviewed by the chief marker. This process is very important in ensuring the integrity of the practical marks.

The predicted mark is the school assessed mark assigned to the artwork/s submitted for the practical (production) examination only.

Submission categories

The submission may be one artwork or a collection or suite of artworks linked conceptually or materially. Artwork that does not comply with category size requirements or is dangerous to handle will be referred to the breach of examination rules committee.

Maximum size, weight or time requirements
Candidates make their submission in one of the three categories listed below. For the purposes of fairness and equity the following requirements regarding the maximum size, weight or time of submitted artwork/s must be adhered to. See Appendix 4: Submission dimensions.

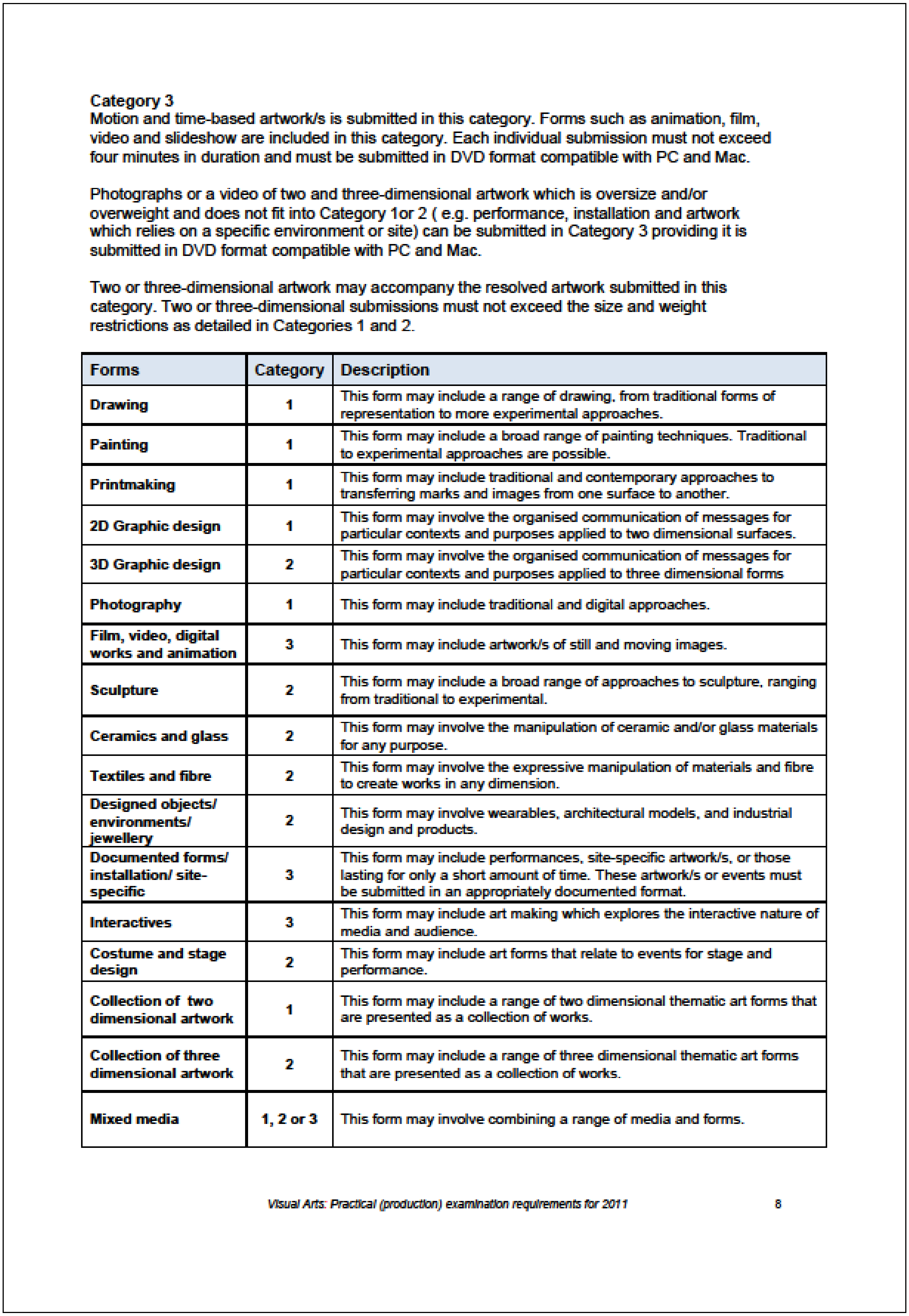
Category 1
Two-dimensional artwork/s is submitted in this category. The complete submission must not exceed two and a half square metres (2.5sqm) when displayed for marking. A single two-dimensional artwork must not exceed 20 kg in weight when packed for marking.

OR

Category 2
Three-dimensional artwork is submitted in this category. The complete submission must not exceed 1.5 cubic metres in volume. Each three-dimensional-artwork must not exceed 20 kg in weight when packed for marking.

Two-dimensional artwork may accompany the resolved artwork submitted in this category. The two-dimensional submission must not exceed the size and weight restrictions as detailed in Category 1.

OR

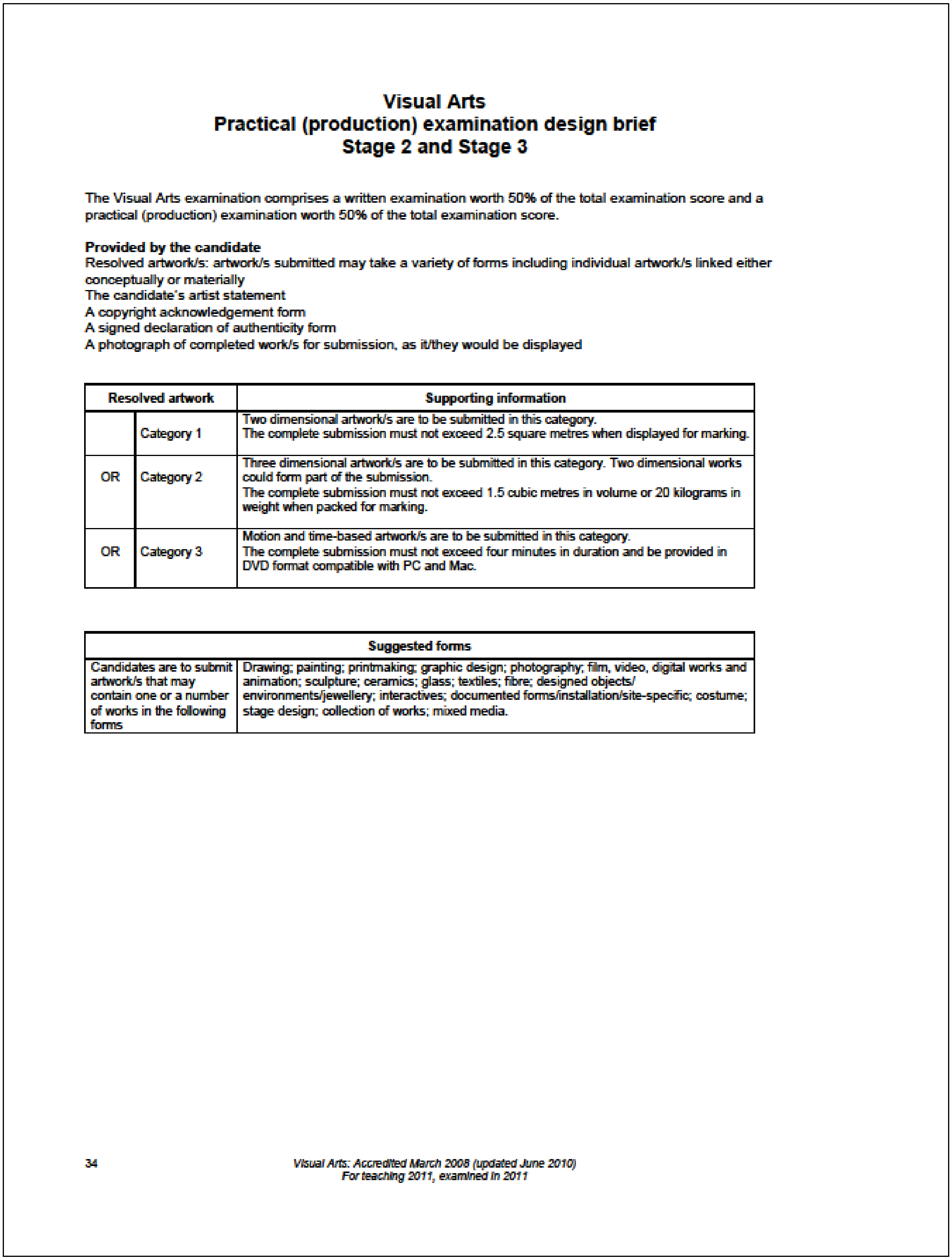


Category 3
Motion and time-based artwork/s is submitted in this category. Forms such as animation, film, video and slideshow are included in this category. Each individual submission must not exceed four minutes in duration and must be submitted in DVD format compatible with PC and Mac.

Photographs or a video of two and three-dimensional artwork which is oversize and/or overweight and does not fit into Category 1or 2 (e.g. performance, installation and artwork which relies on a specific environment or site) can be submitted in Category 3 providing it is submitted in DVD format compatible with PC and Mac.

Two or three-dimensional artwork may accompany the resolved artwork submitted in this category. Two or three-dimensional submissions must not exceed the size and weight restrictions as detailed in Categories 1 and 2.

Forms	Category	Description
Drawing	**1**	This form may include a range of drawing, from traditional forms of representation to more experimental approaches.
Painting	**1**	This form may include a broad range of painting techniques. Traditional to experimental approaches are possible.
Printmaking	**1**	This form may include traditional and contemporary approaches to transferring marks and images from one surface to another.
2D Graphic design	**1**	This form may involve the organised communication of messages for particular contexts and purposes applied to two dimensional surfaces.
3D Graphic design	**2**	This form may involve the organised communication of messages for particular contexts and purposes applied to three dimensional forms
Photography	**1**	This form may include traditional and digital approaches.
Film, video, digital works and animation	**3**	This form may include artwork/s of still and moving images.
Sculpture	**2**	This form may include a broad range of approaches to sculpture, ranging from traditional to experimental.
Ceramics and glass	**2**	This form may involve the manipulation of ceramic and/or glass materials for any purpose.
Textiles and fibre	**2**	This form may involve the expressive manipulation of materials and fibre to create works in any dimension.
Designed objects/ environments/ jewellery	**2**	This form may involve wearables, architectural models, and industrial design and products.
Documented forms/ installation/ site-specific	**3**	This form may include performances, site-specific artwork/s, or those lasting for only a short amount of time. These artwork/s or events must be submitted in an appropriately documented format.
Interactives	**3**	This form may include art making which explores the interactive nature of media and audience.
Costume and stage design	**2**	This form may include art forms that relate to events for stage and performance.
Collection of two dimensional artwork	**1**	This form may include a range of two dimensional thematic art forms that are presented as a collection of works.
Collection of three dimensional artwork	**2**	This form may include a range of three dimensional thematic art forms that are presented as a collection of works.
Mixed media	**1, 2 or 3**	This form may involve combining a range of media and forms.



Visual Arts
Practical (production) examination design brief
Stage 2 and Stage 3

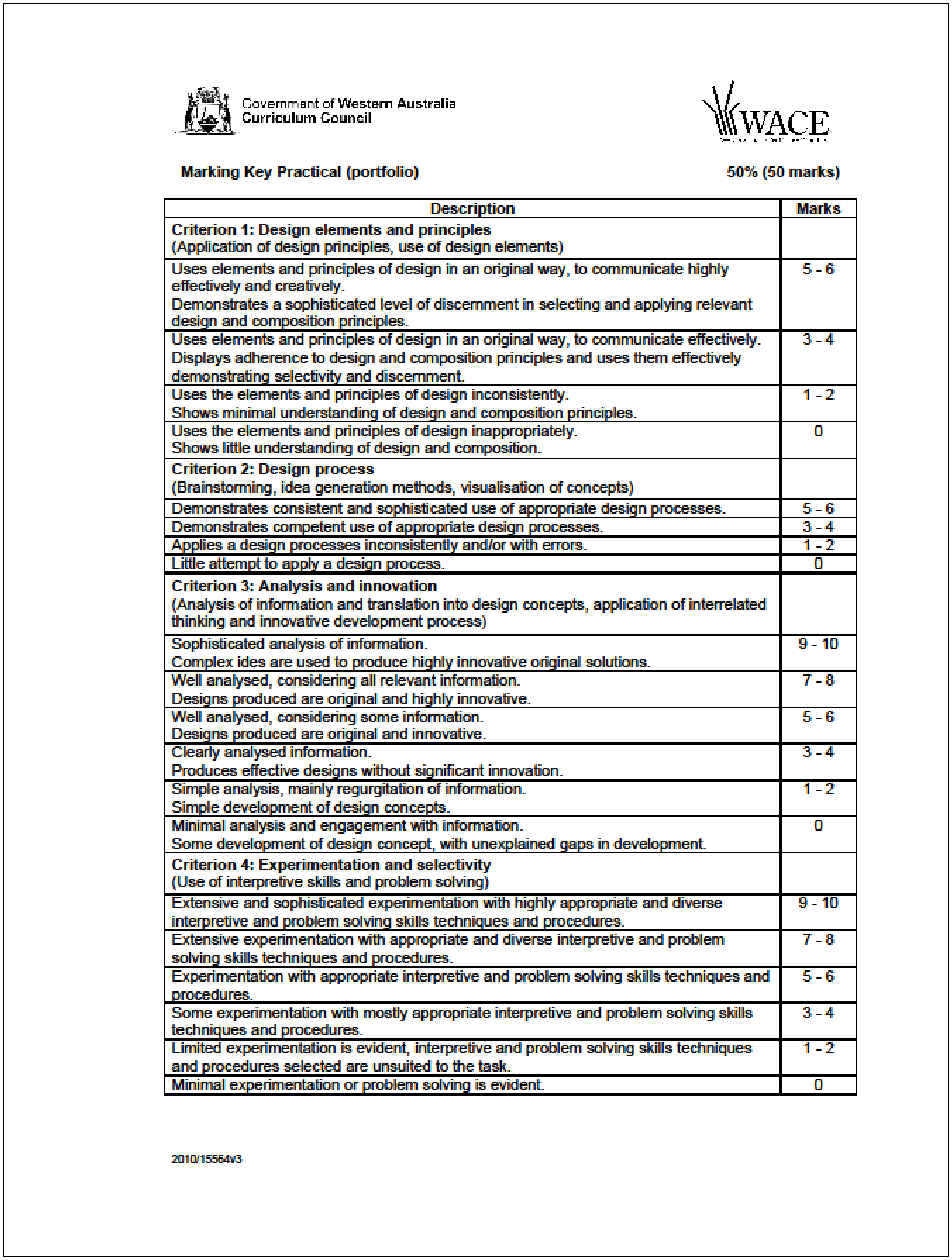
The Visual Arts examination comprises a written examination worth 50% of the total examination score and a practical (production) examination worth 50% of the total examination score.

Provided by the candidate
Resolved artwork/s: artwork/s submitted may take a variety of forms including individual artwork/s linked either conceptually or materially
The candidate's artist statement
A copyright acknowledgement form
A signed declaration of authenticity form
A photograph of completed work/s for submission, as it/they would be displayed

Resolved artwork		**Supporting information**
	Category 1	Two dimensional artwork/s are to be submitted in this category. The complete submission must not exceed 2.5 square metres when displayed for marking.
OR	Category 2	Three dimensional artwork/s are to be submitted in this category. Two dimensional works could form part of the submission. The complete submission must not exceed 1.5 cubic metres in volume or 20 kilograms in weight when packed for marking.
OR	Category 3	Motion and time-based artwork/s are to be submitted in this category. The complete submission must not exceed four minutes in duration and be provided in DVD format compatible with PC and Mac.

Suggested forms	
Candidates are to submit artwork/s that may contain one or a number of works in the following forms	Drawing; painting; printmaking; graphic design; photography; film, video, digital works and animation; sculpture; ceramics; glass; textiles; fibre; designed objects/ environments/jewellery; interactives; documented forms/installation/site-specific; costume; stage design; collection of works; mixed media.

Appendix C Marking Key – Design



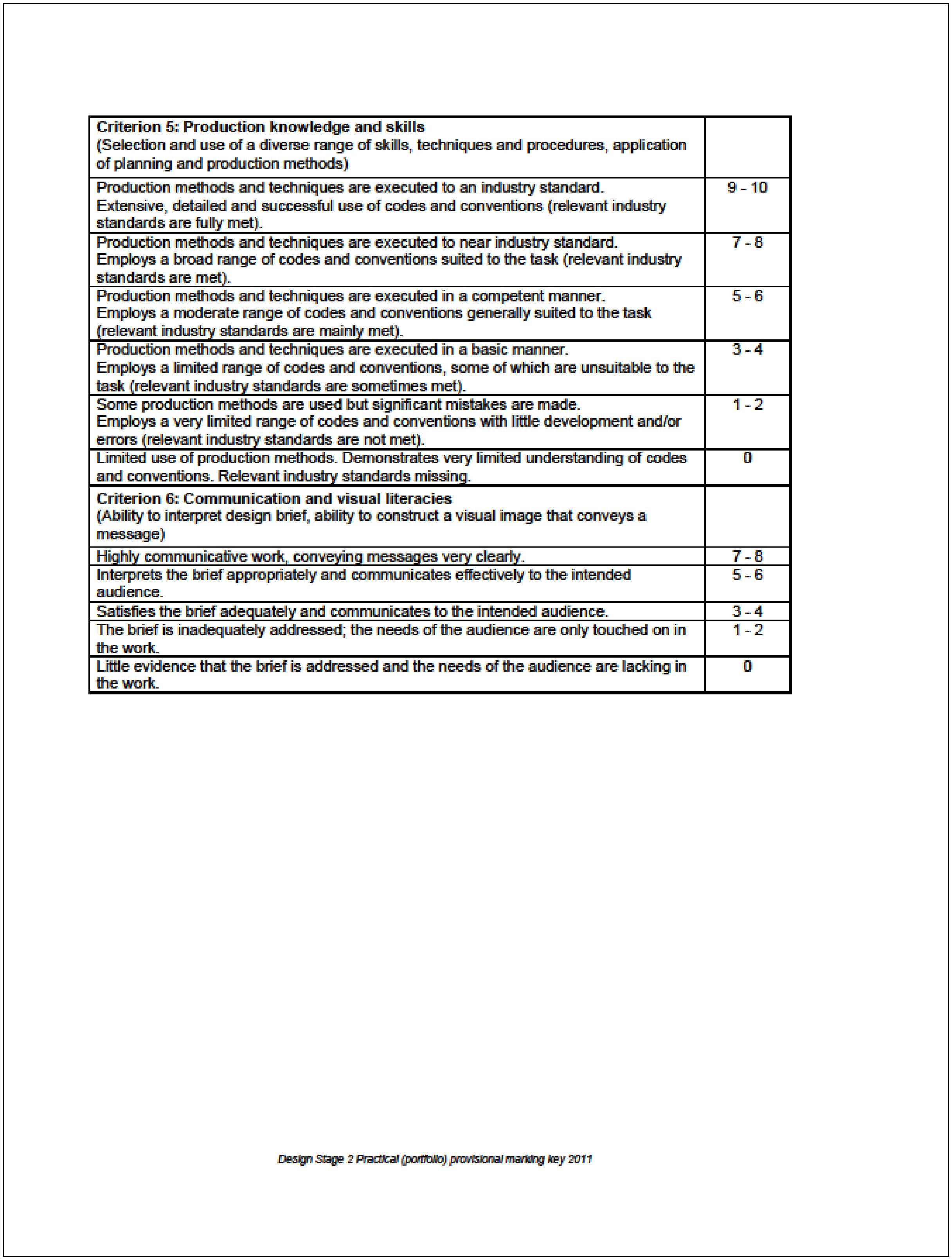
Government of Western Australia
Curriculum Council

WACE

Marking Key Practical (portfolio) **50% (50 marks)**

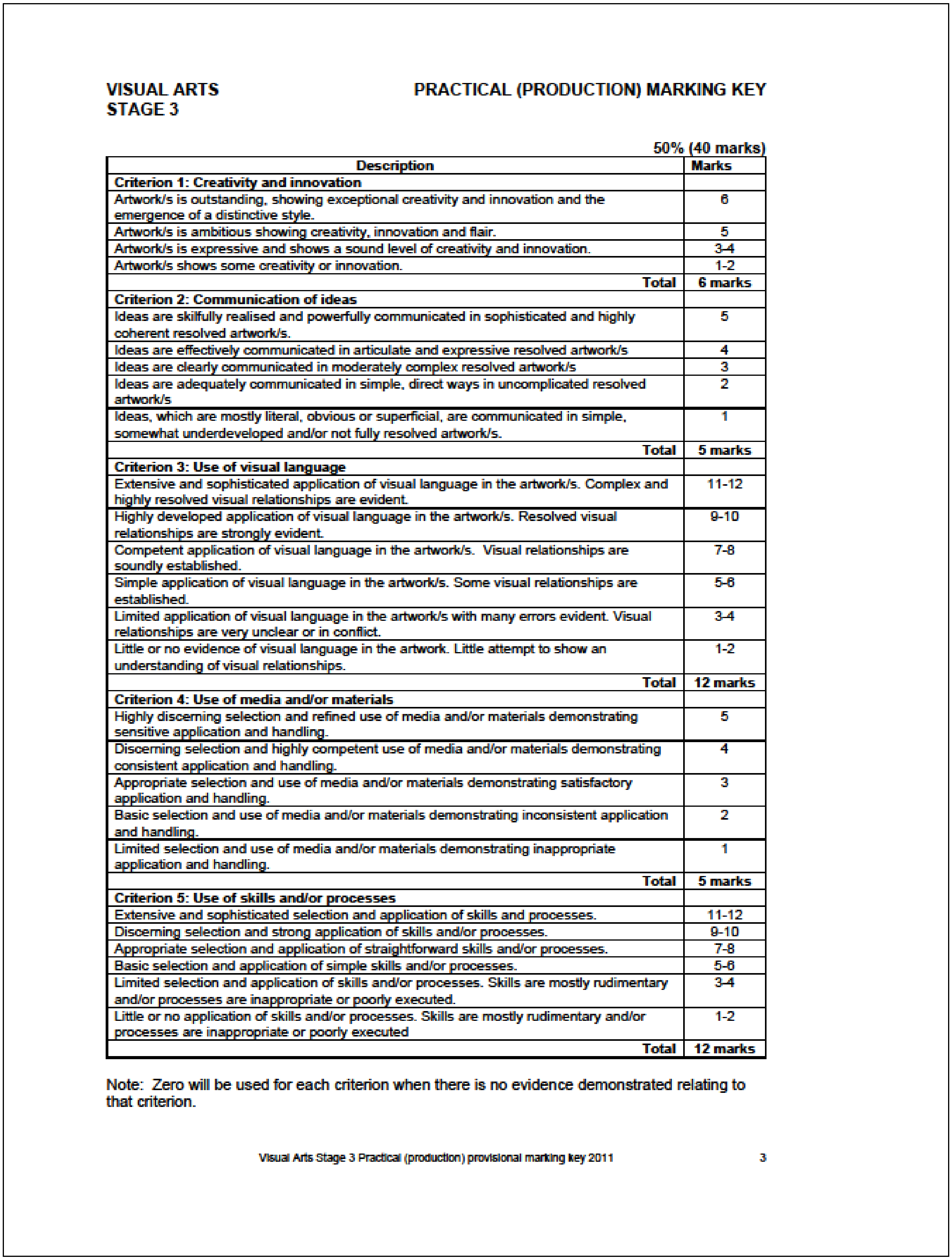
Description	Marks
Criterion 1: Design elements and principles (Application of design principles, use of design elements)	
Uses elements and principles of design in an original way, to communicate highly effectively and creatively. Demonstrates a sophisticated level of discernment in selecting and applying relevant design and composition principles.	5 - 6
Uses elements and principles of design in an original way, to communicate effectively. Displays adherence to design and composition principles and uses them effectively demonstrating selectivity and discernment.	3 - 4
Uses the elements and principles of design inconsistently. Shows minimal understanding of design and composition principles.	1 - 2
Uses the elements and principles of design inappropriately. Shows little understanding of design and composition.	0
Criterion 2: Design process (Brainstorming, idea generation methods, visualisation of concepts)	
Demonstrates consistent and sophisticated use of appropriate design processes.	5 - 6
Demonstrates competent use of appropriate design processes.	3 - 4
Applies a design processes inconsistently and/or with errors.	1 - 2
Little attempt to apply a design process.	0
Criterion 3: Analysis and innovation (Analysis of information and translation into design concepts, application of interrelated thinking and innovative development process)	
Sophisticated analysis of information. Complex ides are used to produce highly innovative original solutions.	9 - 10
Well analysed, considering all relevant information. Designs produced are original and highly innovative.	7 - 8
Well analysed, considering some information. Designs produced are original and innovative.	5 - 6
Clearly analysed information. Produces effective designs without significant innovation.	3 - 4
Simple analysis, mainly regurgitation of information. Simple development of design concepts.	1 - 2
Minimal analysis and engagement with information. Some development of design concept, with unexplained gaps in development.	0
Criterion 4: Experimentation and selectivity (Use of interpretive skills and problem solving)	
Extensive and sophisticated experimentation with highly appropriate and diverse interpretive and problem solving skills techniques and procedures.	9 - 10
Extensive experimentation with appropriate and diverse interpretive and problem solving skills techniques and procedures.	7 - 8
Experimentation with appropriate interpretive and problem solving skills techniques and procedures.	5 - 6
Some experimentation with mostly appropriate interpretive and problem solving skills techniques and procedures.	3 - 4
Limited experimentation is evident, interpretive and problem solving skills techniques and procedures selected are unsuited to the task.	1 - 2
Minimal experimentation or problem solving is evident.	0

2010/15564v3



Criterion 5: Production knowledge and skills (Selection and use of a diverse range of skills, techniques and procedures, application of planning and production methods)	
Production methods and techniques are executed to an industry standard. Extensive, detailed and successful use of codes and conventions (relevant industry standards are fully met).	9 - 10
Production methods and techniques are executed to near industry standard. Employs a broad range of codes and conventions suited to the task (relevant industry standards are met).	7 - 8
Production methods and techniques are executed in a competent manner. Employs a moderate range of codes and conventions generally suited to the task (relevant industry standards are mainly met).	5 - 6
Production methods and techniques are executed in a basic manner. Employs a limited range of codes and conventions, some of which are unsuitable to the task (relevant industry standards are sometimes met).	3 - 4
Some production methods are used but significant mistakes are made. Employs a very limited range of codes and conventions with little development and/or errors (relevant industry standards are not met).	1 - 2
Limited use of production methods. Demonstrates very limited understanding of codes and conventions. Relevant industry standards missing.	0
Criterion 6: Communication and visual literacies (Ability to interpret design brief, ability to construct a visual image that conveys a message)	
Highly communicative work, conveying messages very clearly.	7 - 8
Interprets the brief appropriately and communicates effectively to the intended audience.	5 - 6
Satisfies the brief adequately and communicates to the intended audience.	3 - 4
The brief is inadequately addressed; the needs of the audience are only touched on in the work.	1 - 2
Little evidence that the brief is addressed and the needs of the audience are lacking in the work.	0

Appendix D Marking Key – Visual Arts



VISUAL ARTS
STAGE 3

PRACTICAL (PRODUCTION) MARKING KEY

50% (40 marks)

Description	Marks
Criterion 1: Creativity and innovation	
Artwork/s is outstanding, showing exceptional creativity and innovation and the emergence of a distinctive style.	6
Artwork/s is ambitious showing creativity, innovation and flair.	5
Artwork/s is expressive and shows a sound level of creativity and innovation.	3-4
Artwork/s shows some creativity or innovation.	1-2
Total	**6 marks**
Criterion 2: Communication of ideas	
Ideas are skilfully realised and powerfully communicated in sophisticated and highly coherent resolved artwork/s.	5
Ideas are effectively communicated in articulate and expressive resolved artwork/s	4
Ideas are clearly communicated in moderately complex resolved artwork/s	3
Ideas are adequately communicated in simple, direct ways in uncomplicated resolved artwork/s	2
Ideas, which are mostly literal, obvious or superficial, are communicated in simple, somewhat underdeveloped and/or not fully resolved artwork/s.	1
Total	**5 marks**
Criterion 3: Use of visual language	
Extensive and sophisticated application of visual language in the artwork/s. Complex and highly resolved visual relationships are evident.	11-12
Highly developed application of visual language in the artwork/s. Resolved visual relationships are strongly evident.	9-10
Competent application of visual language in the artwork/s. Visual relationships are soundly established.	7-8
Simple application of visual language in the artwork/s. Some visual relationships are established.	5-6
Limited application of visual language in the artwork/s with many errors evident. Visual relationships are very unclear or in conflict.	3-4
Little or no evidence of visual language in the artwork. Little attempt to show an understanding of visual relationships.	1-2
Total	**12 marks**
Criterion 4: Use of media and/or materials	
Highly discerning selection and refined use of media and/or materials demonstrating sensitive application and handling.	5
Discerning selection and highly competent use of media and/or materials demonstrating consistent application and handling.	4
Appropriate selection and use of media and/or materials demonstrating satisfactory application and handling.	3
Basic selection and use of media and/or materials demonstrating inconsistent application and handling.	2
Limited selection and use of media and/or materials demonstrating inappropriate application and handling.	1
Total	**5 marks**
Criterion 5: Use of skills and/or processes	
Extensive and sophisticated selection and application of skills and processes.	11-12
Discerning selection and strong application of skills and/or processes.	9-10
Appropriate selection and application of straightforward skills and/or processes.	7-8
Basic selection and application of simple skills and/or processes.	5-6
Limited selection and application of skills and/or processes. Skills are mostly rudimentary and/or processes are inappropriate or poorly executed.	3-4
Little or no application of skills and/or processes. Skills are mostly rudimentary and/or processes are inappropriate or poorly executed	1-2
Total	**12 marks**

Note: Zero will be used for each criterion when there is no evidence demonstrated relating to that criterion.

Visual Arts Stage 3 Practical (production) provisional marking key 2011 3

Appendix E Assessor Interview – Design

Assessor Questions

<u>Digital Portfolios - Design</u>

A. Background

1. In 2011 you were ……………… years old.
2. You have ………………..years of teaching experience, in which …………………..years have been in Design.
3. Please list your qualifications.

General	Related to Design

4. In WACE 2011, did you (please circle):

   a. Teach Stage 2 Design students? Yes / No

   b. Mark the real Stage 2 Design student works? Yes / No

B. Comments

Please make comments under the following headings:

Digital representations: authenticity and quality

- suitability of digital representation for course
- breadth does the digital representation allow all students to demonstrate performance
- limitations of the digital representation

Comparative pairs marking process and online tool

- ease of accessing student work

- ease of entering judgements

- ease of making judgements

- suggestions for improvements

Quality of student work

- general standard of work

- factors that may have influenced standard of work

- opportunity for students to demonstrate quality

Any other comments:

Appendix F Assessor Interview – Visual Arts

Assessor Questions

<u>Digital Portfolios – Visual Arts</u>

A. Background

1. In 2011 you were ……………… years old.
2. You have ………………..years of teaching experience, in which ……………………years have been in Visual Arts.
3. Please list your qualifications.

General	Related to Visual Arts

4. In WACE 2011, did you:

   a. Teach Stage 2 Visual Arts students? Yes / No

   b. Mark the real Stage 2 Visual Arts student works? Yes / No

B. Comments

Please make comments under the following headings:

Digital representations: authenticity and quality

- suitability of digital representation for course

- breadth does the digital representation allow all students to demonstrate performance

- limitations of the digital representation

- types of the digital representation necessary (pdf files, picture files, videos, etc.)

Comparative pairs marking process and online tool

- ease of accessing student work

- ease of entering judgements

- ease of making judgements

- suggestions for improvements

Quality of student work

- general standard of work

- factors that may have influenced standard of work

- opportunity for students to demonstrate quality

Any other comments: