The viability of simulated large-scale marking as professional development for preservice teachers

Nathanael Reinertsen

Follow this and additional works at: https://ro.ecu.edu.au/theses

Part of the Educational Assessment, Evaluation, and Research Commons, and the Teacher Education and Professional Development Commons

This Thesis is posted at Research Online.
Edith Cowan University

Copyright Warning

You may print or download ONE copy of this document for the purpose of your own research or study.

The University does not authorize you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site.

You are reminded of the following:

• Copyright owners are entitled to take legal action against persons who infringe their copyright.

• A reproduction of material that is protected by copyright may be a copyright infringement. Where the reproduction of such material is done without attribution of authorship, with false attribution of authorship or the authorship is treated in a derogatory manner, this may be a breach of the author’s moral rights contained in Part IX of the Copyright Act 1968 (Cth).

• Courts have the power to impose a wide range of civil and criminal sanctions for infringement of copyright, infringement of moral rights and other offences under the Copyright Act 1968 (Cth). Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.
The viability of simulated large-scale marking as professional development for preservice teachers

This thesis is presented for the degree of

Doctor of Philosophy

Nathanael Reinertsen

Edith Cowan University
School of Education
2020
Abstract
Judging the quality of student work is a core skill of a proficient teacher. This professional competency is often utilised by organisations that run large-scale marking operations when they recruit teachers as markers. These organisations and the teachers themselves often claim that large-scale marking is valuable professional development.

This research aimed to determine whether professional learning outcomes similar to those reported by experienced teachers can be achieved for preservice teachers through participation in a live simulation of a large-scale marking operation. The research was conducted in three phases: an online survey of Australian teachers to establish that reports of benefit from other contexts are generalisable to Australia; a simulated marking experience for 22 preservice teachers at Edith Cowan University; and finally, follow-up interviews with seven of those participants after their first semester of teaching to establish to what extent their perceptions of the marking experience had changed.

The research collected qualitative data through interviews and an online survey. Additionally, there was quantitative data collected during the marking simulation in the form of scores, and these were analysed with simple descriptive statistics, linear correlation, and a many-facet Rasch analysis to examine the severity or harshness of the novice markers. The analysis of the scoring was not done to examine the skill of the markers, but rather to evaluate the quality of the marker training in the simulation. It was found that the markers scored reliably, and so it was inferred that the simulation training was probably similar to authentic marker training.

The research found that the benefits described by the simulation participants largely centred on increased confidence in marking and gaining experience of marking. These main benefits and several minor ones broadly aligned with benefits published in the literature. The perception of value in the simulated marking experience did not diminish after the preservice teachers had begun their work as teachers, and several reported using processes or concepts from the experience in their professional work.

The research concluded that simulated marking sessions have applications in preservice teacher education. There were strong recommendations from participants that a practical marking experience such as the simulation become a mandatory part of initial teacher education courses.
Copyright 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;

Nathanael Reinertsen
Acknowledgements
Sometimes, it is important to state the obvious. There would be no thesis without my family, my supervisors, the organisation that employed me throughout the writing of this thesis and my colleagues there, and the participants in this research.

My parents, The Reverends Jon and Rae Reinertsen, have always been effusively proud and supportive. I learned from a very young age to love reading and asking questions, and they made many sacrifices for me and for my education. My sister Hannah, too, has always been ready with sisterly advice and love. Without such a beginning, I doubt any of my life’s work would have been possible.

The beginning of my PhD journey was the work of my loving wife (before she was my wife), Esmeralda. Always frank and honest with me, she told me in no uncertain terms to stop dreaming of going back to study and instead to do it! She has always been generous and understanding in giving me the time to work, and I hope that she has not felt abandoned during the evenings I spent writing. I love you, Es.

The idea for this research project, and the corpus of student writing marked in it, sprang from my work at the Australian Council for Educational Research (ACER). I cannot overstate how much the experiences and opportunities I have had at ACER contributed to the skills I needed to complete this research. Special notes of thanks are due to Professor Geoff Masters AO, CEO for his permission to use ACER data in this research; Dr Annie Brown and Dr Tom Lumley for inspiring me; Dr Dan Cloney for troubleshooting my ConQuest code; Louise Courtney for her unwavering support and encouragement; and Kat McGill for her collegiality and friendship.

My principal supervisor, Dr Bill Allen, was a font of encouragement and support. His patience and pure skill as a teacher and mentor were instrumental in every part of this research. A/Prof Brian Moon gave invaluable advice and direction as needed, and Dr Alistair Campbell provided his technical skill in the construction of the marking interface.

The twenty-two preservice teachers who participated in the simulation in this research, and the anonymous survey respondents, generously volunteered their time and without them there would be no data to report.

This thesis was copyedited in accordance with the Institute of Professional Editors Guidelines for editing research theses (2019) by Dr Jennifer Butler, who I also thank sincerely.
In the final months of work on this thesis, I became a father for the first time. So, while I acknowledge that I owe the completion of this research to all the people above, and more whom I have undoubtedly forgotten to name, I dedicate this thesis to my daughter.

To Ada Esther Reinertsen.
# Table of Contents

List of Tables .................................................................................................................................................. x  
List of Figures .................................................................................................................................................... xi  
Definition of Terms and List of Acronyms ........................................................................................................... xii  

**Chapter 1 – Introduction to the Research Project** ......................................................................................... 1  
The Research Problem ........................................................................................................................................ 2  
The Aims of the Research .................................................................................................................................. 5  
The Reasons for the Research ............................................................................................................................ 6  
The Structure of This Thesis ............................................................................................................................... 8  

**Chapter 2 – Literature Review** ...................................................................................................................... 10  
Assessing Writing .................................................................................................................................................. 10  
What Teachers Learn from Participating in Large-scale Marking .................................................................... 12  
Researching Teacher Professional Development ............................................................................................... 14  
Effective Professional Development for Teachers .............................................................................................. 15  
The Role of Developing Identity in Assessment ................................................................................................. 17  
Practice Architectures ...................................................................................................................................... 20  
Simulation in Education ..................................................................................................................................... 20  
Microteaching .................................................................................................................................................... 22  
Simulation in nursing education ......................................................................................................................... 23  
Simulating a Writing Marking Experience .......................................................................................................... 24  
Assessment preparation in Initial Teacher Education ....................................................................................... 25  
Chapter Summary ............................................................................................................................................... 25  

**Chapter 3 – Methodology and Design** .......................................................................................................... 27  
Pragmatism 27  
Practice Architectures ......................................................................................................................................... 30  
Setting and Context ........................................................................................................................................... 32  
Design 33  
Data Analysis ......................................................................................................................................................... 34  
Ethical Considerations ....................................................................................................................................... 36  
Informed consent ................................................................................................................................................ 36  
Anonymity and confidentiality ............................................................................................................................ 37  
Safety .................................................................................................................................................................. 38  
Chapter Summary ............................................................................................................................................... 39
### Chapter 4 – Phase One: Survey

#### Methods

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survey Platform</td>
<td>41</td>
</tr>
<tr>
<td>Sampling</td>
<td>41</td>
</tr>
<tr>
<td>Participant Recruitment</td>
<td>41</td>
</tr>
<tr>
<td>Questionnaire Design</td>
<td>42</td>
</tr>
<tr>
<td>Questionnaire data analysis</td>
<td>44</td>
</tr>
</tbody>
</table>

#### Results

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Characteristics of Respondents</td>
<td>45</td>
</tr>
<tr>
<td>Agreement with Statements</td>
<td>47</td>
</tr>
<tr>
<td>Reasons for Becoming a Marker</td>
<td>49</td>
</tr>
<tr>
<td>Expectations of Marking</td>
<td>55</td>
</tr>
<tr>
<td>Advice to Potential Markers</td>
<td>58</td>
</tr>
<tr>
<td>Barriers to Marking</td>
<td>64</td>
</tr>
</tbody>
</table>

#### Discussion

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clarifying goals and expectations</td>
<td>69</td>
</tr>
<tr>
<td>Deepening discipline knowledge</td>
<td>70</td>
</tr>
<tr>
<td>Learning more about students and their work</td>
<td>70</td>
</tr>
<tr>
<td>Developing insights that support professional practice</td>
<td>71</td>
</tr>
<tr>
<td>Minor themes</td>
<td>72</td>
</tr>
<tr>
<td>Barriers and drawbacks</td>
<td>73</td>
</tr>
</tbody>
</table>

#### Chapter Summary

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>75</td>
</tr>
</tbody>
</table>

### Chapter 5 – Phase Two: Simulation

#### Methods

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants</td>
<td>78</td>
</tr>
<tr>
<td>Researcher’s Role</td>
<td>80</td>
</tr>
<tr>
<td>Data Collection</td>
<td>80</td>
</tr>
<tr>
<td>Time and Location</td>
<td>81</td>
</tr>
<tr>
<td>Student Writing</td>
<td>81</td>
</tr>
<tr>
<td>Marking Procedures</td>
<td>82</td>
</tr>
<tr>
<td>Marking Design</td>
<td>82</td>
</tr>
<tr>
<td>Data Analysis</td>
<td>84</td>
</tr>
<tr>
<td>Interrater Agreement</td>
<td>85</td>
</tr>
<tr>
<td>Investigating Rater Severity with Rasch Analysis</td>
<td>87</td>
</tr>
</tbody>
</table>
Results 89

Quantitative Analysis of Interrater Reliability ................................................... 90
Rasch Analysis of Rater Severity ......................................................................... 92
Limitations of the Quantitative Analyses ............................................................. 96
Qualitative Analysis of the Researcher’s Perspective ........................................... 96
Discussion 100

Chapter Summary .................................................................................................. 104

Chapter 6 – Participant Experiences of the Marking simulation ......................... 105

Methods 105

Data Collection ........................................................................................................ 105
Data Analysis ........................................................................................................... 107

Results 108

Pre-Simulation Interviews and Email Responses .................................................. 109
Summary of key themes on the topic of marking experience ............................... 113
Goals for participating ............................................................................................. 113
Summary of key goals for participating .................................................................. 118
Summary of Themes in Pre-Simulation Interviews and Emails .......................... 118
Post-Simulation Interviews and Emails ................................................................. 119
Summary of Themes in Post-Simulation Interviews and Emails .......................... 133

Discussion 135

What participants report about the simulated marking experience .................... 135
Participants’ reports and teachers’ experiences of large-scale marking ............... 137

Chapter Summary .................................................................................................. 139

Chapter 7 – Phase Three: Follow-up ................................................................. 140

Methods 141

Participants ............................................................................................................. 141
Data Collection ....................................................................................................... 141
Data Analysis .......................................................................................................... 142

Results 143

Experiences of assessment .................................................................................... 144
Initial Teacher Education ....................................................................................... 146
Relevance of the simulation to professional practice ........................................... 149
Outcomes ................................................................................................................. 150
Future marking opportunities .............................................................................. 152
Feedback for future iterations ................................................................. 152
Recommendations for in-service professional development .................. 154
Other positive aspects of the simulation .............................................. 155
Discussion 155
Chapter Summary ................................................................................. 157

Chapter 8 – Synthesis and Discussion .................................................... 159
Summary of Results of Phase One .......................................................... 160
Summary of Results of Phase Two .......................................................... 161
Summary of Results of Phase Three ......................................................... 163
Synthesis 163
The Benefits of Large-scale Marking for Teachers and Preservice Teachers ..... 164
Fidelity, reality and simulations ............................................................... 165
Practical Marking Experience and Initial Teacher Education .................... 168
Simulated Large-scale Marking as Professional Development for Preservice Teachers ........................................................................... 171

Chapter 9 – Conclusion ............................................................................ 173
Limitations 175
Significance 175
Implications for Future Research ............................................................. 176
Conclusion 178
References 179

Appendices
Appendix A: Information Letters ............................................................. 190
Appendix B: Consent Forms ..................................................................... 196
Appendix C: Phase One Questionnaire Items .......................................... 199
Appendix D: Phase One Codebook ............................................................. 201
Appendix E: Simulation Marking Interface ................................................. 203
Appendix F: Proforma Marking Feedback Letter to Simulation Participants ....... 204
Appendix G: ConQuest Command Files for Rasch Analysis ...................... 205
Appendix H: Phase Two Interview Schedules .............................................. 206
Appendix I: Phase Three Interview Schedule ............................................. 209
List of Tables

Table 1 Teaching area specialisations of respondents.......................................................... 46
Table 2 Statements about the effects of participating in large-scale marking. ........ 48
Table 3 Codes for ‘Reasons for marking.’ ................................................................. 50
Table 4 Sub-codes within ‘Professional development.’ ................................................. 50
Table 5 Codes for ‘Expectations for marking.’ .......................................................... 55
Table 6 Codes for ‘Advice to potential markers.’ ....................................................... 59
Table 7 Sub-codes within ‘Professional development benefits.’ ................................. 61
Table 8 Codes for ‘Barriers to marking.’ ..................................................................... 65
Table 9 Participants’ major teaching specialisations by course............................... 80
Table 10 Marking bundle assignments and percentage complete......................... 84
Table 11 Marking summary statistics........................................................................... 91
Table 12 Themes in pre-simulation interviews and emails..................................... 109
Table 13 Themes in post-simulation group interviews and emails........................... 120
Table 14 Codes ordered by number of references..................................................... 144
Table 15 Summary of guiding questions relevant to each chapter and phase........ 160
Table 16 Central tasks of learning to teach. From Feiman-Nemser (2001)............. 170
List of Figures

Figure 1 The media and spaces in which sayings, doings, and relatings exist. From Kemmis et al. (2014, p. 34) ................................................................. 32
Figure 2 Structure of the research. .......................................................................................................................... 34
Figure 3 The Interactive Model. Cited in Miles, Huberman & Saldaña (2014) ........ 35
Figure 4 Questionnaire design .................................................................................................................................. 43
Figure 5 Agreement calculations spreadsheet for Marker_01. ........................................ 86
Figure 6 Analysis 1: Rater severity estimates ................................................................................................. 92
Figure 7 Analysis 2: rater severity estimates ................................................................................................. 93
Figure 8 Analysis 2: Wright map displaying person, rater and criteria locations ...... 95
Figure 9 Extract from the researcher’s field notes. ..................................................................................... 97
Figure 10 Comparison of cases and nodes in the small-group interviews.................... 121
Figure 11 Categories and subcategories of critical components, from Stains & Vickrey (2017, p. 4). ................................................................................................. 177
## Definition of Terms and List of Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACER</td>
<td>Australian Council for Educational Research</td>
</tr>
<tr>
<td>AITSL</td>
<td>Australian Institute for Teaching and School Leadership</td>
</tr>
<tr>
<td>ECU</td>
<td>Edith Cowan University, Western Australia</td>
</tr>
<tr>
<td>ECU HREC</td>
<td>Edith Cowan University Human Research Ethics Committee</td>
</tr>
<tr>
<td>ITE</td>
<td>Initial Teacher Education</td>
</tr>
<tr>
<td>Large-scale marking</td>
<td>A scoring operation for an assessment at a regional, state, national or international level, typically administered to thousands of candidates.</td>
</tr>
<tr>
<td>NAPLAN</td>
<td>National Assessment Program – Literacy and Numeracy (Australia’s national school assessment)</td>
</tr>
<tr>
<td>NESA</td>
<td>New South Wales Education Standards Authority</td>
</tr>
<tr>
<td>PA</td>
<td>Practice Architectures</td>
</tr>
<tr>
<td>PD</td>
<td>Professional Development</td>
</tr>
<tr>
<td>Preservice Teacher</td>
<td>A university student undertaking an initial teacher education course</td>
</tr>
<tr>
<td>SCSA</td>
<td>School Curriculum and Standards Authority (Western Australia)</td>
</tr>
<tr>
<td>TRBWA</td>
<td>Teacher Registration Board of Western Australia</td>
</tr>
<tr>
<td>VCAA</td>
<td>Victorian Curriculum and Assessment Authority</td>
</tr>
<tr>
<td>VCE</td>
<td>Victorian Certificate of Education</td>
</tr>
<tr>
<td>WACE</td>
<td>Western Australian Certificate of Education</td>
</tr>
</tbody>
</table>
Chapter 1 – Introduction to the Research Project

The ability to make reliable judgments about the quality of student work is a core skill of a proficient teacher (Australian Institute for Teaching and School Leadership (AITSL), 2011). This professional competency is often exploited in large-scale marking operations that seek to use teachers with relevant experience as markers. One common justification for why teachers should apply to be markers is that it is valuable professional development (PD). While the data on what teachers learn from large-scale marking are mainly anecdotal and self-reported, many markers and researchers appear to accept that there are benefits from participating in marking.

Often, opportunities to participate in large-scale marking are reserved for experienced teachers, as there is a belief that time spent teaching curriculum content is associated with a deeper knowledge of the assessment construct which leads to greater reliability in judging the quality of student work. This means that what is seen as a valuable PD opportunity for teachers is not often extended to early career teachers. Therefore, this research aims to determine whether similar professional learning outcomes can be achieved for preservice teachers from a simulated marking session.

A preliminary research question, to be refined further in this chapter, was “To what extent can participation in a large-scale marking simulation provide a professional development experience similar to authentic large-scale marking?” If it can be established that a simulation of large-scale marking provides a similar experience to an authentic marking session, simulated marking sessions could have important applications for preservice teacher education, and as a method of providing PD for early career teachers who are still refining their professional judgment.

The idea for this research arose in the course of the researcher’s professional life. Throughout his teaching career, he encountered colleagues who articulated similar beliefs about the benefits teachers derive from participating in exam marking. He continues to encounter similar beliefs about these professional benefits in his current employment at the Australian Council for Educational Research, where a part of his role during the research presented in this thesis was training and supervising markers. In looking to the literature for an exploration and explanation of the grounds upon which these beliefs have been built over time, he found that there were more questions than answers. Goldberg (2012) made a point that resonated deeply: “We must seek to supplement our beliefs and the predominantly anecdotal data on the impact on teachers of involvement in scoring by conducting research to
augment our knowledge and address many still-unanswered questions” (p. 45). This call to elucidate the beliefs and perceived benefits of participation in marking was the beginning of the research reported here.

The Research Problem

As the Australian Professional Standards for Teachers (Australian Institute for Teaching and School Leadership (AITSL), 2011) acknowledge, the ability to make reliable judgments about the quality of student work is a core skill of a proficient teacher. This is evident particularly in Standard 5, ‘Assess, provide feedback and report on student learning,’ Standard 5.1, ‘Assess student learning’ and Standard 5.3, ‘Make consistent and comparable judgments’ (Australian Institute for Teaching and School Leadership (AITSL), 2011). This professional competency of experienced teachers is recognised and often exploited by large-scale marking operations that seek to use teachers with relevant experience as markers.

Curriculum and standards authorities, in Australia and overseas, that recruit markers for exam marking often make claims about how marking benefits teachers, and present these claims as incentives over and above the financial remuneration markers receive. For example, the Victorian Curriculum and Assessment Authority (VCAA) says that “Assessing VCE [Victorian Certificate of Education] external assessments provides a valuable professional development and learning experience” (Victorian Curriculum and Assessment Authority (VCAA), 2018). Similarly, one of the examination boards in the UK, the Assessment and Qualifications Alliance (AQA), has produced a marker recruitment video that repeats some of the common assertions as to why teachers should take up marking opportunities (AQA, 2019). The benefits of marking given in the video include monetary compensation but also the opportunity to see a wide range of student work and network with colleagues.

Accompanying the video on the AQA website is a list of benefits that includes “boost your career, CV and personal development.” The AQA is just one exam board in the UK, and they are used here not as an authority for such claims, but rather as one example of where a common, underlying belief about marking experience being beneficial for teachers has been made explicit and public.

Around the time that this research was being conceptualised and designed, The Age, a newspaper published in Victoria, Australia, described teachers who mark the VCE exams as “the secret weapons helping schools achieve impressive VCE results” (Cook & Butt, 2017). That article reported on the number of teachers at prestigious schools in Victoria who are VCE assessors, equating high numbers of exam markers in a school with success in the VCE
exams. The article quoted Haileybury School principal Derek Scott as saying, “you get a better understanding of what the marking process is and what they are looking for and you can adjust your classroom practice” (Cook & Butt, 2017). This statement draws a direct link from marking experience to improved classroom practice. An anonymous VCAA spokesman is also quoted in the article, asserting that “there is a ‘general association’ between high-achieving schools and larger numbers of assessors” (Cook & Butt, 2017). These statements hint at a pervasive belief that participation in large-scale marking is beneficial for teachers and their students.

In Western Australia, it is mainly teachers who mark the written parts of the annual Western Australian Certificate of Education (WACE) exams, just as it is in Victoria with the VCE. The School Curriculum and Standards Authority (SCSA) is the authority that recruits markers for WACE, the National Assessment Program – Literacy and Numeracy (NAPLAN) and the Online Literacy and Numeracy Assessment (OLNA). SCSA does not make the same public claims about the benefits of marking in their recruitment materials. Anecdotally, Western Australian teachers hold beliefs similar to those of their colleagues in other states and countries about the benefits of participating in large-scale marking (B. Allen, personal communication, February 15, 2017). One of the first objectives of this research was to investigate the extent to which Australian teachers in general hold these beliefs, and as is reported in due course, it was found that teachers do see large-scale marking as ‘good PD.’

What teachers learn from participating in large-scale marking has been the subject of research in a number of contexts, mostly in North America (Falk & Ort, 1998; Gambell & Hunter, 2004; Goldberg, 2012; Goldberg & Roswell, 2000) but with at least one study closer to home in New Zealand (Gilmore, 2002). These studies relied largely on self-reports from the participants collected through interviews and conversations, and the researchers analysed the responses for common themes. The effects of marking experiences on teachers identified in these studies are covered in more detail in Chapter 2. Generally, though, all the participants in the various studies maintained that there was at least some benefit gained from the experience in terms of professional knowledge.

Despite the general belief held by many teachers and researchers that marking is beneficial for teachers, not all teachers have access to the opportunity to participate in large-scale marking and gain whatever benefits there may be. When this research began, the first selection criterion the SCSA had for WACE exam markers was that the applicant needed “recent experience teaching the course at Year 12 level” (Schools Curriculum and Standards
SIMULATED LARGE-SCALE MARKING

Authority (SCSA), 2016). That has since changed to read, “recent or current experience in teaching the course being examined or a similar course” (Schools Curriculum and Standards Authority (SCSA), 2019). Despite the change, there is still an assumption in this selection criterion that experience delivering curriculum content is directly related to a deeper understanding of the assessment construct, and that using experienced teachers increases the reliability of the marking. Similar requirements of experience are required in other jurisdictions. The need for prima facie validity in high stakes assessments is indisputable, and using experienced teachers to mark the exams is one measure that lends support to the validity of the testing program. A drawback of this recruitment criterion is that early career teachers, who may be still in the early stages of developing their assessment skills, may be denied the opportunity to benefit from marking.

Some providers offer professional learning opportunities that emulate some aspects of large-scale marking. For example, the New South Wales Education Standards Authority (NESA) offers to NSW-registered teachers two 12-hour online courses on how to interpret and apply the NAPLAN writing marking criteria (New South Wales Education Standards Authority (NESA), 2019). NESA links this learning to ATISL standards 5.3 and 6.3 (Australian Institute for Teaching and School Leadership (AITSL), 2011). But the opportunity is not intended for beginning or early career teachers: NESA states on their website that completing these courses contributes to “maintaining Proficient Teacher Accreditation in NSW” (New South Wales Education Standards Authority (NESA), 2019). ‘Proficient Teacher’ is a reference to one of the levels of teacher proficiency in the AITSL Professional Standards (Australian Institute for Teaching and School Leadership (AITSL), 2011), above ‘Graduate,’ which is the level associated with early career teachers.

So, there is a situation where teachers are encouraged to gain large-scale marking experience for a range of reasons, but where teachers early in their careers are not afforded the opportunity to participate in authentic large-scale marking because of their perceived lack of expertise and/or experience. Additionally, where there exist PD courses that centre on large-scale marking, beginning teachers are not intended to be the teachers who participate. Assuming, perhaps naively, that authentic marking yields the most authentic benefit, and drawing a conclusion that beginning teachers are denied these opportunities because of concerns about their reliability due to inexperience and the face validity of the assessment, the research problem can be stated as: How can we offer early career teachers the opportunity to learn from the experience of large-scale marking without compromising assessment validity?
The Aims of the Research

This research aimed to evaluate a possible way of delivering to preservice teachers the experience of participating in large-scale marking by simulating a large-scale marking session. Running a simulated large-scale marking experience as targeted PD is not a novel idea (Bejar, 2012). However, there is currently no published example of the idea being put into practice. There has been research into whether preservice teachers’ knowledge, assessment skill and self-efficacy can be improved through structured marking practice through a web-based resource (Dempsey, PytlíkZillig, & Bruning, 2009); this research has some parallels to that work, but its aims are significantly different. The present research does not attempt to quantify the magnitude or strength of particular, hypothesised benefits, but rather aims to identify the range of impacts reported by preservice teachers who have participated in a simulation of large-scale marking, and to compare that data with the impacts and beliefs reported by teachers in a broader context.

The simulation invited preservice teachers who had not participated in large-scale marking before to mark a corpus of previously marked scripts in an environment that was representative of authentic large-scale marking operations, including the marker training that always precedes such marking operations. The markers had an opportunity to form a very temporary community of professionals who collaborate and learn from one another through conversations and common experience – this is not far from the definition of a community of practice: “Communities of practice are 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.” (Wenger-Trayner & Wenger-Trayner, 2015, p. 1) The obvious difference is the temporary nature of the association, compared to the regular interactions of a Community of Practice.

The simulation differed from a real marking operation in minor but significant ways. Firstly, the scripts had all been scored and the results already reported, so there was no impact upon test-takers if an inexperienced marker’s professional judgment was poor or inconsistent. Secondly, the duration of the simulation was far shorter than an authentic marking operation: the simulation was a single day. Thirdly, the corpus of pre-scored scripts was curated such that it included, in its entirety, the full range of quality of student performances, ensuring to the extent possible that the participants were exposed to a range of quality. Exposure to a wide and varied range of student performances is posited by some to be one of the aspects of marking that helps support a teacher’s ability to judge student work (Gambell & Hunter, 2004; Goldberg & Roswell, 2000; Masters & Forster, 2000).
In short, the aim of this research was to establish whether a simulation of a large-scale marking session could deliver to preservice teachers an experience similar to that of experienced teachers who have participated in large-scale marking. The central question for this research was “To what extent can participation in a large-scale marking simulation provide, for preservice teachers, a PD experience similar to authentic large-scale marking?”

The guiding questions for this research were:
1. What research exists about the benefits of participation in large-scale marking?
2. What benefits do experienced Australian teachers report they have gained from participation in large-scale marking?
3. What are the opportunity costs, difficulties or other impacts reported by Australian teachers about their participation in large-scale marking?
4. Can large-scale marking be effectively simulated?
5. What do participants in a simulation of large-scale marking report about the experience?
6. Is the experience reported by participants in the simulation of large-scale marking similar to the experience reported by experienced teachers who have participated in authentic marking operations?
7. After having transitioned from preservice to early career, will the participants’ impressions of the experience have changed compared to those captured immediately after the marking session?

There were three phases of research conducted in order to answer these questions. In phase one, responses from Australian teachers about their participation in marking were collected through an online survey and the results were compared to themes present in the research literature. Phase two was to conduct a simulation of large-scale marking for final-year preservice teachers, capturing information about their perceptions, experiences and opinions of the simulation through observation and semi-structured interviews before and after the simulation. Phase three was to interview participants almost 12 months later, after their first semester of teaching, to see whether their perception of the value of the experience had changed with time and with some teaching experience.

The Reasons for the Research

The process of becoming a marker rather than a reader is an aspect of teaching that may or may not be formally taught in teacher education courses, and yet we expect proficient teachers to be able to make reliable judgments about the quality of a student’s ideas.
expressed in writing in many subject disciplines, not only in English. The absence of evidence about how preservice teachers are prepared to be assessors in Australian universities has been noted (Grainger & Adie, 2014). The assumption that beginning teachers learn these skills in the workplace through department moderation sessions and conversations with colleagues is not unreasonable, but the quality of such on-the-job training will vary considerably depending on context. This research explored another way of supporting teachers to develop their professional judgment while they are still enrolled in a structured teacher-training course.

As already noted, the reason experienced teachers are recruited in authentic large-scale testing is for the sake of validity and reliability. It is assumed that their professional judgment has already been sufficiently developed to a level where their judgments are reliable. Potentially compromising the validity of an assessment program by using preservice teachers just to see what they might learn is inadvisable, at best. Simulation offers a way to deliver the experience of large-scale marking, without the risk of invalidating the results.

In a simulation of marking, the marking guides, processes and student writing can be authentic, without needing to report the marks to students or teachers. The marking session can be as life-like as possible, with the notable difference that there are no high or low stakes: there are no stakes at all. This means that getting the marks ‘right’ is only important for the markers, not for the students who created the work.

Simulation offers several other advantages that make it suitable as a method for this research. One obvious advantage of simulated marking over authentic marking is the ability to tailor the marking session to its participants with regards to time, location and duration. The flexibility is convenient, but other advantages are more substantial.

How to apply a rubric can be taught in online courses (such as the NESA course previously mentioned) or in printed literature. Some assessment organisations, such as Educational Testing Service, train markers online and have them mark from home (Educational Testing Service (ETS), 2019). However, there is a suggestion in the literature that one of the key sources of any benefits that may come from marking is conversations and interactions with teaching colleagues (see Chapter 2 and Masters & Forster, 2000). Simulation, in the context of this research, offered a physical location and a collaborative task where participants were encouraged to share with their peers and the researcher their interpretations of the quality of the texts with respect to several criteria.

In addition to marking their own class’ work, the other type of marking most teachers participate in is in-school moderation. Large-scale marking, whether simulated or not, is
SIMULATED LARGE-SCALE MARKING

distinct from in-school moderation in significant ways. In the first place, it involves marking the work of students about whom a rater has no other information. When marking the work produced by his or her own class, or by students in his or her school, a teacher has knowledge about what has been taught and some degree of knowledge about how a student has performed in the past and other factors that may have affected that student’s performance. Markers of large-scale assessments have no other knowledge than the task requirements, the marking guide and the script in front of them. This forces markers to focus on the quality of the artefact as the only basis for judgment (Wyatt-Smith & Castleton, 2005). For those new to such marking, this may be a novel experience.

Another difference is that marking in an environment where time and marking rates are monitored may have an effect on how a marking guide is applied. The need for fast but accurate judgments may be a novel experience for new markers, too, particularly if their marking experience has yet to include working to tight deadlines.

Lastly, the diversity of student responses and approaches is quite likely to be greater than within a single classroom or even a single school. Authentic large-scale marking requires a high rate of reading and judging scripts. To replicate this condition, a simulation must expose markers to a high volume of scripts, and thus a larger range of student responses than teachers are likely to encounter on a practicum or in a classroom.

This research explored a new way to support the PD of preservice teachers. It did so in a ‘sandbox,’ where preservice teachers had an experience that was comparable to authentic large-scale marking, but without the potential for invalidating the assessment, or ‘getting it wrong’ for the students whose writing was scored.

The Structure of This Thesis

This research comprised three distinct but connected phases of data collection. This thesis, therefore, has been structured to support understanding about how each of these phases is unique, yet related to the others. In Chapter 2, the literature around the topic is examined. Then there is a methodology chapter, Chapter 3, which gives the theoretical framework for the research and describes the principal methods used. After this, the phases are treated and reported separately. Chapter 4, on the survey phase (phase one), details the methods of the online survey, reports the results, then discusses them. The next two chapters present more detail about the simulation phase (phase two), again outlining the method before presenting results and discussion: Chapter 5 gives the researcher’s perspective, then Chapter 6 reports the participants’ perspectives. In Chapter 7, the follow-up phase (phase three) is
SIMULATED LARGE-SCALE MARKING

described. The method for conducting the interviews is outlined, then the results are presented and discussed. Chapter 8 synthesises the results from the three phases, and Chapter 9 presents an answer to the research question.

It is perhaps unconventional to have descriptions of methods included in many chapters rather than placing them together in a single chapter. It is hoped, though, that the reading experience of this presentation will allow for a consideration of each phase as semi-independent, building upon the work presented previously.
Chapter 2 – Literature Review

This chapter provides a review of literature related to this research. The literature reviewed is organised into seven topics: assessing writing; what teachers report learning from large-scale writing assessment; research into PD; the characteristics of effective PD; the role of developing identity in assessment; simulation in education; and simulating a marking experience. The aim of this chapter is to provide some background to the choices made in designing this research. It also aims to establish some of the understandings that are later used to evaluate the central activity of this research: the marking simulation. Even though the six topics covered in this review may appear to be eclectic, each is relevant to one or more important aspect of the marking simulation. Where pertinent, the relevance of the literature discussed in each section is pointed out, although detailed discussion of the research is left to other chapters.

Assessing Writing

Writing is a complex sociocultural phenomenon. It entails not just cognitive processes, but navigating social and cultural conventions (Weigle, 2002). Reading involves interpreting a text, taking into account, to a greater or lesser extent, the context of its production, which may include the writer’s identity and personal history (Wyatt-Smith & Castleton, 2005). Assessing a piece of writing is not only reading it, but also coming to some conclusion or judgment about its quality or merits (Weigle, 2002). What constitutes quality or merit in a piece of writing depends on the purpose of the assessment. An assessment may aim to provide diagnostic information about a student’s writing, and have very explicit criteria (Knoch, 2009, 2011). Alternatively, the assessment domain may be defined more broadly, with the goal of the assessment being to assess some aspect of a student’s skill or knowledge in a more general sense (Weigle, 2002). The selection of analytic or holistic marking is a further aspect of an assessment that is guided by its purpose and whether the assessment domain is defined narrowly or broadly. A decision must be made whether to create an analytic scoring framework with explicit criteria describing and defining quality, or whether to require markers to make an on-balance judgment of the strengths of a response weighed against its weaknesses to arrive at an overall judgment of quality (Knoch, 2009, 2011; Weigle, 2002). Even though analytic scoring and holistic scoring are often treated as two separate and distinct types of marking, each of these two categories can be further divided (Knoch, 2009), indicating that perhaps there is a continuum rather than a dichotomy.
As noted in Chapter 1, classroom teachers assessing the writing of students who are known to them read and judge their students’ writing differently to how they read and judge writing when they have no information about the identity or personal history of the writer and text (Wyatt-Smith & Castleton, 2005). When reading a student’s text, the teacher is able to interpret stylistic choices knowing what instruction the child has received. The teacher is also able to link particular authorial choices to feedback they provided the student in class. When this contextual information is unavailable to teachers, their marking becomes more focused on the artefact, which stands as the only evidence upon which teachers can assess the student’s ability (Wyatt-Smith & Castleton, 2005).

When the piece of writing in front of them is the sole basis for making a judgment of quality, the processes by which raters arrive at their judgments are uncertain, despite attention in language testing research (Bejar, 2012; Eckes, 2008). Studies of rater cognition have shown that there are differences in how novice and experienced teachers assess some aspects of writing (Cumming, 1990), and even when given training and explicit criteria, raters can have idiosyncratic understandings of those criteria (Lumley, 2002). In short, how raters understand and respond to a text and then assign scores according to the meaning they have made, remains unclear.

Even though it is poorly understood how markers arrive at their judgments, it is still a priority in writing assessment to ensure, to the greatest extent possible, that the same piece of writing will receive similar marks from different markers (Brown, Glasswell, & Harland, 2004). In practice, almost all educational and psychological testing examines and reports measures of interrater reliability as a way of examining scoring accuracy (Stemler, 2004). Writing assessment is no different from other educational and psychological testing in this regard, with large-scale writing assessments using a range of statistical measures of interrater reliability to gauge scoring accuracy (Brown et al., 2004). There are three main types of metrics used to gauge interrater reliability: consensus, consistency, and measurement (Stemler, 2004). Consensus is often estimated through percentage exact and adjacent agreement, and consistency is reported through correlational statistics (Pearson’s $r$ most frequently, although Cronbach’s alpha is reported where it can be used appropriately). Measurement estimates are calculated through factor analysis methods, or, since Item-Response Theory has come to the fore in educational measurement, the many-faceted Rasch model (MFRM) (Stemler, 2004).

However interrater reliability is estimated, certain rules-of-thumb for acceptable rates of interrater agreement can be inferred from examining the literature on large-scale tests of
SIMULATED LARGE-SCALE MARKING

writing, as Brown et al. (2004) did in their work validating a New Zealand writing assessment’s marking guide. They found that consensus estimates in the studies reviewed for exact agreement range typically between 40% and 60% and adjacent agreement rates of 80%–100%, are commonly found. Consistency coefficients of between .70 and .80 are reasonably consistently reported for standardized performance assessments … Measurement coefficients, reported much less often in the studies, tend to fall in the range .60 to .80. (p. 110)

These general limits of acceptable accuracy reported in the literature indicate that despite the complexity of marking as a task and the idiosyncrasies of markers’ judgment processes, broad agreement is not only possible, but expected in large-scale marking.

What Teachers Learn from Participating in Large-scale Marking

The PD benefits teachers gain from participating in large-scale marking have been investigated in many contexts, but the data are mainly anecdotal or self-reported through interviews. This qualitative approach – interviewing teacher-markers and recording their reflections about where they felt they had benefitted – has been pursued by several of the researchers in this field (Darling-Hammond & Falk, 2013; Falk & Ort, 1998; Gambell & Hunter, 2004; Gilmore, 2002; Goldberg & Roswell, 2000). Additionally, there has been a recent research study into teachers’ motivations, beliefs and values around participating in large-scale assessment as a form of PD, in the US (Palermo & Thomson, 2019).

The research thus far has found that teachers see participation in large-scale marking in a positive light, and report gaining awareness of students’ literacy behaviours and confirming their confidence in their ability to judge the quality of student work (Gambell & Hunter, 2004). Other reported benefits include validation of teachers’ expertise and a “greater understanding of teaching and assessing practices and processes” (Gilmore, 2002) as well as teachers becoming “more reflective, critical and deliberate” (Goldberg & Roswell, 2000). Reported intrinsic motivations for participating include professional growth, better supporting students’ learning outcomes, and refining assessment practice (Palermo & Thomson, 2019).

Falk and Ort (1998) identified four areas in which scoring provided teachers with professional learning: clarifying their goals and expectations; deepening their discipline knowledge; learning more about students and their work; and developing insights that support their professional practice. Additionally, they reported that the marking experience strengthened teachers’ sense of professionalism. Darling-Hammond and Falk (2013) built on this earlier work to argue for greater utilisation of scoring as PD, asserting benefits such as deepening teachers’ understandings of skills contained in curricula, reinforcing common
standards across jurisdictions, building communities of practice, shifting testing culture in schools, and ultimately improving teaching and learning. All these benefits are alluded to in other research, including in Falk and Ort (1998). Palermo and Thomson (2019) reported that their findings continued to align with the findings of Falk and Ort (1998).

In contrast to the four main areas Falk and Ort (1998) identified, Goldberg (2012) asserted that “What teachers report learning from scoring has tended to center around the assessment itself …, rather than on broader implications for instructional practice” (p. 44). Such contrasting views are not uncommon in the literature on this topic, which indicates the lack of consensus as to what teachers learn from participating in marking. The probable reason for such contradictory views is that there is a lack of evidence to suggest that the knowledge and skills an individual gains from a marking session can be generalised, not just across different assessment programs, but across all participants in the same marking session. For example, Goldberg (2012) went on to state that “Until now the assertion that scoring serves (or can serve) as professional development has tended to get passed along without considering the great variety of experiences through which teacher engagement in scoring is filtered” (p. 44). The experience of scoring, and the benefits derived from it, may be quite personal or may differ from one experience to the next. The benefits gained by one participant may not be shared by all participants, and the benefits gained from one style or type of marking may not be shared among different marking designs or contexts. Even how each marker in a marking process interprets, understands and internalises the marking criteria seems to be more personal and complex than is evident from commonly used statistical measures of consistency. As noted in the previous section, investigation of markers’ understandings of criteria has been attempted through studying rater cognition (Bejar, 2012; Wyatt-Smith & Castleton, 2005), and each marker’s understanding and application of criteria appears to vary considerably and idiosyncratically (Lumley, 2002).

One recognised gap in the literature is investigation into how the benefits of participation in marking translate into classroom practice (Goldberg, 2012). For all the self-reported value of the experience, there has yet to be an investigation into how growth in a teacher’s ability to ‘Make consistent and comparable judgments’ (Australian Institute for Teaching and School Leadership (AITSL), 2011) directly impacts on student learning. It still remains to be proven that teachers who have participated in marking are more effective in the classroom and that their students progress any differently, in terms of rate or level of achievement, in their learning.
Nevertheless, that there are benefits from participation in marking and that teachers do learn from the experience seems to be well accepted by all participants and researchers, even if the precise nature of those benefits seems still to be unclear. Whatever the gains, the reported benefits of participation appear to have two main sources: the first is conversations with colleagues about student work, and the second is exposure to a wide range of student work of varying quality (Gambell & Hunter, 2004; Goldberg & Roswell, 2000; Masters & Forster, 2000). Whether one or the other of these contributes more to professional knowledge is difficult to establish. It is still common practice that large-scale marking operations, particularly those that use criterion-based marking guides, require markers to meet (whether in person or online) to ensure they apply the marking guide consistently, and these meetings involve discussion and close reading of student writing. In these situations, marking cannot be done without conversations about student writing, making these two aspects of marking practice interdependent.

**Researching Teacher Professional Development**

It has already been noted that what teachers learn from large-scale marking is somewhat contested (Goldberg, 2012). Additionally, there is an apparent gap in the literature about how Australian preservice teachers are taught to become assessors (Grainger & Adie, 2014). The present research aimed to address both of these deficiencies by researching a specific instance of a PD activity; in order to meet this aim, it is appropriate to understand the nature of researching PD interventions. Borko (2004) presented a three-phase framework for researching and understanding PD activities for teachers. In the first phase, the researcher focuses on a single instance of a program at a single site, and aims to collect “existence proofs of effective professional development.” Borko described research in this phase as being centrally concerned with the relationship between the PD program and teachers, and noted that such research is likely to be labour intensive; be conducted by the program designer; and involve volunteers who are interested in innovative ideas. Her second phase shifts focus to the implementation of a single PD program at multiple sites with multiple facilitators, and the focus of her third phase is broader still, comparatively evaluating a range of PD programs run at many sites by many facilitators.

The research reported in this thesis fits very well in Borko’s first phase, as it is research into ‘existential proofs’ that simulated large-scale marking can be effective as PD, specifically for preservice teachers. This is a useful characterisation, and it is used in Chapter 9, where the research project and its contribution are considered as a whole. There are several
aspects to these ‘existential proofs’ that must be considered. One important aspect is identifying the difference between effective and ineffective PD activities, and determining what evidence is sufficient to categorise a particular activity as ‘effective.’

**Effective Professional Development for Teachers**

There has been research into how PD activities can best meet the needs of teachers. Borko, Jacobs and Koellner (2010) carried out a survey of contemporary literature in the US, and identified a range of characteristics of effective PD, both with regards to the content of PD and the processes and structures of effective PD. They identified two strong themes in relation to the content of effective PD: that the content should be “situated in practice” and “focused (at least in part) on student learning”. Of particular note in the context of this research, they identified that in three of the reports they surveyed, it was noted that PD activities should “help teachers learn how to elicit and interpret their students’ ideas, examine student work, and use what they learn about students’ ideas and work to inform their instructional decision and actions” (p. 550). The emphasis on examining students’ work and understanding their ideas resonates strongly with the central activity and aim of the present research; evaluating students’ writing is absolutely central to the marking simulation.

As well as examining content, Borko and colleagues also identified processes and structures of highly effective PD (Borko et al., 2010). In the literature they surveyed, there were common themes of active participation and collaborative work. Active participation is consistent with constructivist pedagogies, as is collaboration, and this perspective is evident in their discussion of the theme of active participation. It is possible to consider a marking panel as a Community of Practice (Wenger-Trayner & Wenger-Trayner, 2015), and this is relevant in considering marking as professional learning. The panel becomes, if only for a short time, a group of professionals who collaborate and learn from one another through conversations and common experience. Wenger-Trayner and Wenger-Trayner (2015) describe three characteristics that are fundamental to a Community of Practice: a shared domain of interest, community formed through interactions, activities and discussions, and a practice. For a marking panel, education is the domain, and perhaps a specific subject or skill (literacy, for example) within the broader domain of education, and marking itself is the practice. But what of community? Despite the short duration of a single marking operation, it is not uncommon for markers to be re-employed year after year; in this researcher’s professional work, one annual marking panel has been largely unchanged in its composition for the past four years, with some markers having worked on that panel for seven consecutive
years. This has led to a sense of community among the ‘regulars,’ especially those who also
mark together on other projects for other organisations throughout the year, despite the time
that elapses in between.

An additional strong theme Borko et al. (2010) identified was that effective PD is
situated within a school context: “school based.” The authors noted that “artifacts of practice,
such as videotapes of lessons and student work, can be used to bring the classroom to the PD
setting” (p. 550). Again, this resonates with this research, as the marking simulation is
designed to be an active experience that is focused on examining, both collectively and
individually, authentic student work.

A final characteristic Borko et al. (2010) identified is that effective PD “should
consist of activities that are ongoing and sustainable over time.” (p. 550). They talked of the
practices in the PD becoming “cycles of experimentation and reflection” (p. 550) in order to
effect long-term change in teachers’ practice. This is at odds with the one-day workshop
design of this research project and so it may be that the marking simulation could be less
effective than a more sustained program of PD activity. Furthermore, Aubusson, Ewing and
Hoban (2009) classified teachers’ learning activities into three categories: PD, Continuous
Professional Development (CPD) and Professional Learning. They described PD as a
traditional, one-off workshop or school development day activity delivered by an external
expert. They believed it to be much less effective than other forms of teacher learning as
teachers have no control over its content and there is no structure provided for teachers to link
any learning from the PD to their teaching practice. Their recommended form of teacher
learning is ‘Professional Learning’, which includes practices such as action learning, project-
based learning and problem-based learning. It is not evident how marking student work fits
within this classification system. On the one hand, it is an activity and a project where
teachers must apply their existing body of knowledge and experience to a difficult task, and
this may not be characteristic of the traditional, one-off workshop. On the other hand, there is
no clear, structured way for teachers to apply new knowledge derived from the experience in
their teaching practice, and it is certainly not a ‘long-term approach’ to teacher learning. In
short, then, a simulated marking experience does not have all the characteristics of the most
effective forms of PD, in that it is a time-bound experience that provides little opportunity for
structured ongoing learning, even if the practices, experience and new knowledge from the
experience may persist and inform participants’ practice over a longer term.

In the Australian context, the Australian Charter of Professional Learning for School
Leaders and Teachers (Australian Institute for Teaching and School Leadership (AITSL),
sets out that professional learning ought to be relevant, collaborative, and future focused (pp. 6–7). These three elements have straightforward parallels to the characteristics of effective PD described by Borko et al. (2010). For example, the Charter’s “relevant” (p. 6) is similar to Borko and colleagues’ “situated in practice” (p. 549) and “school based” (p. 550). The Charter does not list Borko and colleagues’ research in its list of “evidence that informed the development of the charter” (p. 9). Taken at face value, then, these are two independently developed frameworks for identifying or designing quality PD activities that overlap considerably in the characteristics they have identified as supporting efficacy. This goes some way to supporting an argument that the characteristics of effective PD have been identified, are understood generally, and are applicable across contexts. As Borko et al. concluded, “Research in this field offers a relatively focused and agreed upon direction for PD, which is consistent with prevailing theories about the nature of learning and cognition” (2010, p. 555).

The Role of Developing Identity in Assessment

Identity is a fundamental part of professional practice. As Ewing and Smith say, “Being is embedded in our practice of doing and, through the doing, as practitioners we continue to become who we are” (2001, p. 16). This interconnectedness of practice and identity is likely to be the reason why Borko (2004) recommends that a ‘situative perspective’ is adopted in researching PD, and that perspective is also known as a sociocultural approach. There is a more recent body of research that also uses sociocultural perspectives to examine the development of assessment competency (Adie, 2012, 2013; Adie, Klenowski, & Wyatt-Smith, 2012; Adie, Stobart, & Cumming, 2019; Charteris & Dargusch, 2018; Cowie, Cooper, & Usher, 2014; Hipkins & Robertson, 2012; Looney, Cumming, van Der Kleij, & Harris, 2018; Wyatt-Smith & Castleton, 2005). Several of these pieces of research examined the practice of assessment moderation in assessment and how it informs the development of identity, or more specifically, assessment identity (Looney et al., 2018).

Identities are social constructs, informing both how we are seen and how we see ourselves. It is in the interaction between ourselves and others that they are formed and reformed. An identity is not fixed: parts, or even whole identities, evolve constantly in response to our actions and interactions with others, and as such an identity is a product of context (Day, Kington, Stobart, & Sammons, 2006; Gee, 2000).

The complexity and malleability of identity makes defining types of identity difficult, but in this research, two ‘types’ of identity are referred to: assessment identity and
professional identity. Assessment identity is used to refer to a teacher’s self-concept as an assessor, constructed not only from one’s knowledge and skills, but also from feelings such as one’s confidence to make judgments of students’ work, and one’s beliefs about assessment and its relationship to teaching and learning (Looney et al., 2018). This definition of assessment identity was an extension of the concept of assessment identity used by Xu and Brown (2016), where a teacher’s assessment identity is a consequence of developing assessment literacy, and in fact is said to be “ultimate goal” (p.158) of the conceptual framework they developed for assessment literacy. That conceptual framework does include affective dimensions among teachers’ conceptions of assessment, and by placing it as the ‘goal’ of assessment literacy, they acknowledge identity as being highly significant in teachers’ assessment practices. However, Looney et al. go further in proposing, “a dynamic and interactive teacher assessment identity constituted by beliefs, feelings, knowledge and skills.” (2018, p.14) that is not simply a consequential ‘product’ of learning about assessment.

Professional identity is used in this research as a more general term to refer to self-concept of oneself as a teacher. It is constructed from one’s professional knowledge and practice, but is also shaped by affective aspects such as confidence and a sense of others’ approval (McCormack, Gore & Thomas, 2006). The reason for using a second term is to acknowledge that assessment is only a part of teaching, and so one’s identity as a teacher will necessarily include broader knowledge, skills, feelings and beliefs in addition to those related to assessment.

Interpersonal interaction is a constant in assessment, because assessment is never a solo pursuit. It will always involve, in the minimal case, an assessor and a person being assessed, a teacher and a student. How students see teachers – for example, as fair or unfair, as harsh or lenient, as an expert or novice in the field of study – is one aspect that informs an assessment identity. Another aspect, though, is teachers’ interactions with colleagues such as in social moderation. The development of the assessment identities of practicing teachers through social moderation practices has been investigated in Australia (Adie, 2013; Adie et al., 2012). The research found that conversations about different points of view concerning the assessment standard and the piece of student work being considered – which is the central practice of moderating judgments – affected teachers’ identities. In the present research, conversations about standards and the quality of writing performances are a key part of marker training, though they are different to other types of moderation in that they are led by a marker trainer. It can be proposed, then, that the conversations that take place within even a
simulated large-scale marking have the potential to shape the participants’ identities as assessors.

Within the context of initial teacher education (ITE), the development of a professional identity has been posited as a “central task of learning to teach” in a framework developed by Feiman-Nemser (2001). In that framework, the tasks of learning to teach are divided chronologically, and the development of a professional identity is categorised as a task for ‘induction’ – a time after ITE when a teacher is beginning work as a professional. A small research project, conducted in Australia by McCormack, Gore and Thomas (2006), found that developing professional identity was one of the more challenging aspects of professional learning for the 16 teachers they tracked through their first year of teaching. In the research at hand, the marking simulation experience was delivered to preservice teachers in the final year of their ITE courses. A point for consideration later, then, is whether participating in the simulated practice of moderation and marking is too early in these teachers’ experiences to be able to affect their assessment identities.

Gee (2000) reflected on identities that are produced by dialogue and discourse as being only one type of identity. A different type of identity is formed through experiences and participation in practices. Gee labelled this perspective on identity as the “affinity perspective”, where people are affiliated with an identity because they are seen to have participated in practices associated with that identity. From this perspective, gaining experience in assessment practices may encourage practitioners to self-identify as assessors, as well as to be identified by others as assessors. The perspectives are not mutually exclusive, as identities may be layered or multifaceted. Indeed, Gee considered yet another perspective on identity that stems from institutional recognition. For professional markers, this layer of identity may manifest in a statement such as ‘I am a NAPLAN marker,’ for example, where the individual feels their identity as an assessor is connected to the specific assessment program in which they practice and hold dialogues with colleagues.

Sociocultural perspectives, and the concepts of professional identity and assessment identity, are useful when researching with the aim of collecting ‘existential proofs’ (Borko, 2004) of the effect of professional practices and PD activities. This is because such perspectives take into account not only the participants in the activities, but the social structures, relationships and interactions, as well as the actions that are involved in these complex activities. The development or reshaping of participants’ assessment identities in an assessment-focused PD activity is a salient effect that ought to be considered when evaluating the effectiveness of the activity, particularly when the activity is intended to provide a
pedagogical experience for preservice teachers, who are at the cusp of entering the profession and the induction stage of learning to teach.

**Practice Architectures**

The sociocultural ontologies of any practice, but PD activities in particular, can be difficult to identify, but one emerging theoretical framework for understanding these aspects of practice is Practice Architecture (PA) theory, which is “a contemporary account of social reality that focusses on practice” (Mahon, Kemmis, Francisco & Lloyd, 2017, p. 2).

‘Practice’ in this context is defined as a cooperative activity that involves and is characterised by “forms of understanding (sayings), modes of action (doings), and ways in which people relate to one another and the world (relatings), that ‘hang together’ in a distinctive project.” The project comprises the actors’ intentions, actions and goals (Rönnerman & Kemmis, 2016).

So, because research into PD that Borko (2004) would categorise as first phase research relies on understanding the relationship between teachers and the PD activity, sociocultural perspectives of identity, as well as sociocultural understanding of sayings, doings, and relatings, offer lenses through which a particular PD program or activity can be examined and understood. There is a connection to be made between a point made previously about the potentially individualised benefits of participating in large-scale marking (Goldberg, 2012), which is that the contexts of marking are likely to mean that generalising from a single marking experience to all marking experiences is not possible. The implication, then, is that understanding the context as well as how the individuals within it relate to that context and each other is essential to understanding the activity.

Chapter 3 discusses this in more detail, but it is worth pointing out here that even though sociocultural perspectives are quite useful in understanding PD, this thesis does not adopt a sociocultural perspective throughout. Rather, it primarily adopts a pragmatic approach. Pragmatism is discussed further in Chapter 3, which also contains a more detailed discussion of PA in relation to this research. Because of its usefulness in understanding PD, the theory of PA is drawn upon later, where useful.

**Simulation in Education**

There are a wide range of activities and experiences that fall into the category of simulation. Any activity or experience that is designed to imitate the key characteristics of an authentic activity or experience can be considered a simulation (Girod & Girod, 2008). From
medical students practicing on dummies to flight simulators used in training airline pilots, simulation is used in a wide range of training and education contexts, and simulation has a similarly long history in teacher education (Girod & Girod, 2008).

The breadth of activities that fall under the umbrella of the term ‘simulation’ has led to the need for more exact definitions. Systematic reviews by Sauvé et al. (Sauvé et al., 2005; Sauvé, Renaud, Kaufman, & Sibomana, 2008) have sought to consolidate the definition and essential attributes of educational simulations, and differentiate educational games from simulations. They found the literature they surveyed supported their initial hypothesised set of essential attributes. The essential attributes of an educational simulation are that it is a model of reality defined as a system that is simplified, dynamic, accurate and valid, and that the experience has an educational objective (Sauvé, Renaud, Kaufman, & Marquis, 2007).

A simplified model of reality is necessary because ‘reality’ is an extremely difficult and complex concept to replicate, and this raises an issue that is often considered in simulations: fidelity. A useful, straightforward definition of the term was provided by Sadler (2010): "Fidelity can be thought of as the extent to which something actually is what it purports to be, and is therefore true to type, concept or label. Fidelity is about the identity of the object concerned and how it is constituted, not about its utility for practical purposes” (p. 728). Although that definition was used by Sadler in a different context, looking at fidelity in assessment, the definition of the term is the same as how the term is used in this research: fidelity is the extent to which a simulation resembles the practice it seeks to reproduce.

A simulation cannot hope to replicate the entirety of a ‘real’ situation; it can only emulate as much as is feasible (Sauvé et al., 2008), so simulations differ in respect to how closely they resemble authentic situations. Consider resuscitation dummies used in first-aid training to teach cardiopulmonary resuscitation. Often, the dummy comprises only a stylised torso and head. The manikin is a simplified representation of a human – simplified by removing all but the parts of the body that are related to the educational objective. This reduces the fidelity of the experience: the first-aider-in-training does not have to deal with clothing, fluids or limbs as they may encounter should they ever need to perform CPR in an authentically. Nonetheless, there is a level of accuracy that must be retained in the simulation in order for it to have validity. Continuing with the example of a CPR dummy, the chest is compressible so that the correct action can be taught and there is a mechanism to make the chest rise to replicate lungs inflating. So, while a simulation simplifies reality, there is a minimal level of fidelity required because “the simpler a model is, the more it runs the risk of distorting the reality under study” (Sauvé et al., 2007, p. 252). Too much distortion of
the reality that a simulation seeks to reproduce compromises the validity and reliability of the experience, which decreases its utility as an instructional tool (Sauvé et al., 2007).

It has been argued that simulations, in particular role-plays, deepen learning experiences in educational settings through the addition of an emotional dimension (Clapper, 2010). Research carried out within Edith Cowan University’s (ECU) teacher education program found that simulation was able to engage students emotionally, and also physically, cognitively and socially (de Jong, Lane, & Sharp, 2012). These axes of engagement are precisely why some educators choose simulation over other pedagogical approaches.

In their discussion of instructional design for complex learning, Van Merriënboer and Kirschner (2018) cited earlier work by Hayes and Singer that divided fidelity into three types: psychological, functional and physical. Psychological fidelity is about simulating emotional, social and cognitive states, much like the simulation carried out by de Jong et al (2012). Functional fidelity is about replicating the actions and activities of a situation, while physical fidelity is, as the name suggest, about recreating the environment in which the authentic experience occurs. Van Merriënboer and Kirschner went on to describe how it may be appropriate to accept differing levels of these three types of fidelity based on the stage of learning, using an example within medical education of starting with paper-based case studies, which have low functional and physical fidelity, and moving on to more complex simulations as the learners progress in their skills and understanding. There is a claim that simulations that are high in all three types of fidelity may be less effective when used with novices because there are too many confounding factors that may distract or detract from the simulation’s pedagogical goals.

Microteaching

Microteaching is one example of simulation in teacher education that has been widely adopted: it was reported to be one of the most common practical activities in teacher education courses aside from teaching practicums (Metcalf, 1995), and it continues to be seen as an integral part of teacher education (Hattie, 2009). The term was coined in the 1960s at Stanford University, where it was used to denote an exercise where learners give a very short lesson that is videotaped, with the purpose of reviewing behavioural aspects of the learner’s lesson delivery and providing feedback (Metcalf, 1995). The original process has been adapted in different contexts, but the term is still used to denote similar exercises (see Hattie, 2009). Even though Metcalf treated microteaching as something separate from simulation, it
SIMULATED LARGE-SCALE MARKING

falls under the broader definitions of simulation given by Girod and Girod (2008) and Sauvé et al. (2007).

As a simulation, microteaching can vary considerably in its fidelity in recreating real classrooms. Some microteaching experiences use school-age learners as pupils for the lessons, while in other circumstances peers play that role (Metcalf, 1995). The reduced length of the ‘lesson,’ the inauthenticity of the environment and the pupils, and the presence of a camera recording the lesson are some of the circumstances that reduce the fidelity of microteaching. Researchers have not ignored the problems associated with this: “As in any simplification of practice required for approximations, the nature of the simplification matters. In microteaching what was stripped away may have been the very aspects of teaching that make it difficult” (Grossman & McDonald, 2008, pp. 190–191). Despite the limitations of microteaching, Metcalf (1995) found that it is moderately effective in helping preservice teachers develop instructional behaviours, but for in-service teachers it was not only moderately effective in helping to develop instructional behaviours, but also affect and knowledge. In other words, microteaching is more effective when the teachers participating have had classroom experience. Hattie (2009) found in his seminal meta-analysis that microteaching is a very effective tool in developing teachers’ instructional skills, giving an effect size of $d = 0.88$.

This is just one example of a type of simulation that is a common and effective tool in both teacher education and PD. Microteaching and role-play are tools suited for developing behavioural and pedagogical skills in teachers. What has yet to be established is what type (or types) of simulation is best suited for developing a teacher’s assessment skills.

**Simulation in nursing education**

As already noted, simulation is used in many fields of education and training. A slight aside into simulation in nursing education is included here because of the body of research that has been deliberately assembled by an organisation in the US, the National League for Nursing (NLN), to support a theoretical framework for simulation in nursing education: the NLN/Jeffries Simulation Framework (see Adamson, 2015; Ravert & McAfooes, 2014). As Sanford (2010) said a few years before these publications, there was a “lack of supporting theory and evidence-based research supporting the use of simulation” and that much of the research base for the use of simulated clinical situations were studies that were “little more than … opinion poll[s]” (p. 1010).
The NLN/Jeffries Simulation Framework is a response to the lack of theoretical basis for simulation in nursing education (LaFond & Van Hulle Vincent, 2013). The framework outlines five constructs (Teacher, Student, Educational Practices, Outcomes and Simulation Design Characteristics) and seeks to provide a common definition of what a simulation is to guide researchers and educators who use simulation in nursing education (Ravert & McAfooes, 2014). The framework is not yet concrete, nor has it yet reached the point of being sufficiently supported with empirical evidence (Adamson, 2015; LaFond & Van Hulle Vincent, 2013).

The reason this aside in to nursing education is warranted, is that thus far, the researcher has not found a documented theoretical framework for the use of simulation in teacher education, despite its history of use (Girod & Girod, 2008). There are parallels with the situation pointed out by Lafond and Vincent (2013), who recognised that the research that made reference to theoretical bases referred to general education theories such as constructivism and experiential learning, rather than a theory specific to simulation. With further research, the NLN/Jefferies Simulation Framework might be generalisable to other fields of education. This is, perhaps unfortunately, outside the scope of this present research, but is noted here as a potential future discussion point.

**Simulating a Writing Marking Experience**

In a discussion of potential uses of automatically scored assessments, Bejar (2012) briefly suggested using a corpus of pre-marked student writing to run simulated marking experiences as targeted PD. That is, to take teachers who have not participated in a large-scale marking program before and have them participate in marking a corpus of genuine student responses that have already been marked. The context of Bejar’s comment was a discussion of the possible uses and the validity implications of automated marking systems, which is far beyond the scope of this present research. Nonetheless, it is important to note where the concept was first encountered.

In the US, there has been research examining whether preservice teachers’ knowledge, assessment skill and self-efficacy can be improved through web-based structured marking practice (Dempsey et al., 2009). The present research has some parallels to that work but there are significant differences. Firstly, that study’s aims were to quantify the magnitude of three hypothesised benefits of participation in a virtual, web-based environment. More significantly, although participants interacted with authentic student writing, the number of pieces of student writing was “relatively few” (Dempsey et al., 2009, p. 58). Nevertheless,
their findings were that structured practice in a virtual environment with authentic student writing was able to produce gains in “knowledge, assessment skills and self-efficacy” (Dempsey et al., 2009, p. 58). Also of particular note is the reported finding that the study may provide support for using analytical marking as the most suitable method of producing gains in assessment skill and self-efficacy.

**Assessment preparation in Initial Teacher Education.** Dempsey et al. (2009) went on to point out that the lack of preservice teacher preparation in writing assessment is due largely to its impracticality: “Offering meaningful practice and feedback requires that instructors provide their students with authentic student writing samples and individual feedback on their assessments, resources not typically available to most instructors” (p. 40). It is difficult to establish whether this is the case globally, locally or even in most universities in the US. In the US and Canada, there seems to have been, at least historically, a lack of teacher training in assessment literacy (Plake cited in Mertler, 2009; Mitton-Kukner & Orr, 2014; Popham, 2011). However, the necessity of including assessment as a central topic in ITE has been explicitly recognised in literature on what ought to be included in ITE curricula (e.g. Bahr & Mellor, 2016; Kosnik & Beck, 2009; Shepard et al., 2005).

In Australian universities, there appears to be little literature on how preservice teachers are supported in developing their assessment literacy generally, and their assessment of student writing specifically (Grainger & Adie, 2014). A similar observation, that there was a lack of research into how preservice teachers are trained in assessment, was made in New Zealand some years ago (Hill, Cowie, Gilmore, & Smith, 2010). However, it is plausible that some type of simulation of writing marking is used; it is a current practice in at least one Western Australian university to have preservice teachers mark samples of genuine student work (Moon, B., personal correspondence, February 2017), and this presumably is a strategy used elsewhere. How widespread similar practices are in other universities and how effective such practices are in supporting preservice teachers in developing their assessment skills are topics that appear to be underreported in the literature.

**Chapter Summary**

This chapter reviewed some of the relevant literature in six areas: assessing writing; what teachers report learning from large-scale writing assessment; research into PD; characteristics of effective PD; simulation in education; and simulating a marking experience. The review found that the process of writing assessment is acknowledged to be complex and multi-faceted, but precisely how markers arrive at their judgments is poorly understood. The
review also found that participants reported large-scale writing assessment to be beneficial in a range of ways, though the precise nature of the benefits may be highly dependent on the individual and the context of the experience. There was found to be some consensus on the characteristics of effective PD, several of which are pertinent to this research, including active participation, collaboration, and focus on examining student work. There was also a brief overview of how research into PD can be conceptualised into three phases, and examined from sociocultural perspectives that can help to understand the complexity of marking practice. Another finding was that use of simulation in education has a long history, and depending on how it is defined, there are many strategies and techniques used in teacher education that can be considered to be simulation. With respect to simulating a marking experience specifically, it was found that others have conceived the idea and there is at least one published example of web-based simulation that aimed to support preservice teachers’ development in assessing writing (Dempsey et al., 2009).
Chapter 3 – Methodology and Design

Having, in Chapter 2, reviewed literature on a range of topics pertinent to this research, such as the assessment of writing, inquiry into PD, assessment identity and the use of simulation in education, this chapter turns to how this present research intended to answer its research questions. This chapter explains the pragmatic perspective taken, and the broad design of this research. It begins with a discussion of the research paradigm, then moves on to the consequential choices of methodology and methods. After providing an overview of the research design, including an explanation of the three phases in which the data were collected and analysed, it ends with consideration of the ethical considerations relevant to the research.

The specific methods and details of the design of each of the three phases of this research are explained in detail in the appropriate, subsequent chapters. This chapter is concerned only with the overall methodology and design of the research.

Pragmatism

The research questions for this research arose from a pragmatic approach rather than a paradigm-driven approach (Punch, 2014), and the methods were selected within the same paradigm. Australian teachers now work in a context where their performance is compared to described professional standards. Proficiency in the AITSL Australian Professional Standards for Teachers (Australian Institute for Teaching and School Leadership (AITSL), 2011) is an expectation of registration boards, government policy makers, as well as employers and school leaders. Using simulated marking as targeted PD for beginning and preservice teachers is one potential intervention that may help teachers develop proficiency and competency, better preparing them for their future careers. Thus, the research questions came from “practical and professional issues and problems” (Punch, 2014) about how to support early career teachers in meeting the prescribed standards, rather than from a ‘top-down’ conceptual or theoretical perspective.

Pragmatism is, in part, a rejection of hard divisions in the ‘paradigm wars’ between qualitative and quantitative research (Tashakkori & Teddlie in Plano Clark & Creswell, 2008). Frequently, pragmatism is used as the guiding philosophy of mixed-methods research because of its inclusiveness; it accommodates a range of ontologies, epistemologies, axiologies and methodologies (Johnson, Onwuegbuzie, de Waal, & Stefurak, 2017). Research that takes a pragmatic approach is less interested in definitions of reality and knowledge than it is in finding answers to research questions, using whatever methods are most appropriate to
collect data from and about the phenomenon it is investigating. However, that is not at all to say that pragmatic research ignores these things. Johnson et al. (2017) looked at the ontology, epistemology, axiology, and approach to inquiry of four pragmatists: Peirce, James, Dewey and Rorty. Each of these pragmatists engaged in philosophical debate about reality, truth and how we can know them; however, “they preferred action to philosophizing, where pragmatism is, in a sense, an anti-philosophy” (Johnson et al., 2017, p. 274). Despite the fact that the present research adopts this ‘anti-philosophy,’ a discussion of the stances taken toward ontology, epistemology, and axiology is needed, before moving to a discussion of the logic behind the selection of its methods.

One of the similarities Johnson et al. (2017) saw in the four pragmatists they examined was that

They emphasized that the starting point of philosophy is lived experience, where experience is not merely sense experience or an intuitive, subjective awareness of mental content but includes a world of people, objects, and events that are encountered and mediated through the cultural, historical, and personal backgrounds which, together, comprise perspective. (p. 274)

This strongly resonates with the present research, and encapsulates its ontology. As previously stated, the research questions arose in the course of the researcher’s professional life. Marking is a lived and experienced sociocultural phenomenon: it is interactions between people contextualised within cultural, historical and personal frameworks. Philosophising about how marking can be real is not necessary for the aims of the present research to be achieved. It is enough to recognise that it is a shared, lived experience for thousands of people, and as such is able to be the object of inquiry.

Epistemologically, the present research intended to answer its research questions. Such an endeavour requires a stance from which truths can be found through inquiry. These truths are “Lowercase ‘t’ truths”: they are working assumptions “given through experience and experimenting” (Johnson & Onwuegbuzie, 2004, p. 18) that solve a current problem or can be applied in a given situation. They are easily discarded if they are no longer useful or applicable, rather than being universal or lasting.

Axiologically, one of the general characteristics of pragmatism is that it “takes an explicitly value-oriented approach to research that is derived from cultural values” (Johnson & Onwuegbuzie, 2004, p. 18). The present research seeks to find ways to give preservice teachers opportunities to participate in a practice that they would not otherwise have access to. As stated previously, large-scale marking is usually reserved for experienced teachers for the sake of the face validity of the marking process. Thus, one of the values entrenched in this
research is inclusivity. It aimed to extend opportunities to participate in a beneficial practice to those who had not previously had access to the practice. Additionally, ITE providers owe a duty to their students to prepare them for the work they take on when they become teachers. That work includes meeting professional standards that include judging student work. This is also relevant to the axiology of this research, because in investigating the viability of a method to support preservice teachers in their development as teachers and assessors, it is motivated by seeking to improve the way things are: it is values-driven.

Pragmatism rejects dualisms and adopts an inclusive logic of ‘both-and’ rather than ‘either-or’ logic (Johnson et al., 2017). This is particularly relevant in the selection of methods within a research project, and has led to pragmatism being associated with mixed-methods research. Mixed-methods research does not recognise particular methods as being restricted to only qualitative or only quantitative research, and so freely combines methods within a research study when it is appropriate and useful to do so. If qualitative and quantitative paradigms are taken to be two ends of a spectrum, then mixed-methods research exists between them, turning the dichotomy into a spectrum (Johnson, Onwuegbuzie, & Turner, 2007). Because of the mixture of methods in this research, it is worth considering where this research might fit on that spectrum.

The present research predominantly employed methods that are typically associated with the qualitative research paradigm: surveys, interviews and the researcher’s observations. These methods were used to gather data about how people experience marking, both authentic and simulated. They were chosen as methods that would provide the best chance of capturing participants’ experiences, attitudes and beliefs (Silverman, 2010) about whether and how marking experiences affected them.

However, parts of the survey questionnaire and the marking simulation produced quantitative data. The closed-response questions in the survey yielded numerical data, and the marking exercise, unsurprisingly, produced marks. These marks or scores represent the complex cognitive process of marker judgment, and as such were used to shed light on some aspects of the marking experience such as consistency, reliability and rater effects through statistical analyses. To exclude these data as a source of information because of ‘either-or’ logic would have diminished the quality of this research. Adopting the ‘both-and’ logic of pragmatism allowed for a richer study.

That said, the quantitative component of this study is quite small in comparison to the qualitative component, both in size and richness. As will be described in detail later, interviews with participants were conducted at three different time points, and there were
both individual and group interviews. This body of qualitative data is larger and much deeper than the scoring data, and far more important for the aims of the research. Although the quantitative components of this research contribute meaningfully to its findings, they are not given equal status.

While this research could, perhaps, be labelled as ‘mixed model,’ which is “mixing qualitative and quantitative approaches within or across the stages of the research process” (Johnson & Onwuegbuzie, 2004, p. 20), it most closely resembles what Johnson, Onwuegbuzie and Turner (2007, p. 124) termed ‘Qualitative Dominant’ research. They located Quantitative Dominant research on the qualitative-quantitative spectrum, noting that it straddles the boundary of what is considered mixed-methods, closer to ‘Pure Qualitative.’ So, although this study is guided by the philosophy of pragmatism and although it mixes qualitative and quantitative data and methods, it does not claim to be a mixed methods study: given the qualitative components are given much greater status, it is a Qualitative Dominant research project.

**Practice Architectures**

While the research questions, design and methods were developed according to pragmatism, some aspects of the discussion and understanding of the simulation that is at the heart of this research necessitated a specific theoretical paradigm. So, the discussion of the findings of this research in Chapter 8 draws on the theory of Practice Architectures (PA) (Kemmis et al., 2014). PA is itself a stand-alone theory situated within Practice Theory (Mahon et al., 2017). It became apparent after data collection and during the analysis of the data that a structured theoretical framework was needed, because while pragmatism was wholly suitable for planning and facilitating the simulation, it offered little guidance for understanding the many complex interactions within the simulation. Such interactions include those between the participants, the rubric and the work being marked; between the participants and the facilitator in forming consensus understandings of the marking guide; and between participants and their peers as they discussed scoring the pieces in front of them. All of these aspects of the simulation are important elements in understanding its impact, and pragmatism alone was not sufficient for making sense of the simulation.

PA offers a way to understand a practice. Within PA, a practice is conceptualised as a project, where people come together to perform an activity to achieve particular goals. This definition is simple, and intuitively applicable to a marking operation, which is where a panel of markers come together to work on a project with the objective of reliably scoring student
work. In this research, though, the project is not an authentic marking operation and the goal is not to produce valid and reliable scores; the goal is to provide a learning experience for preservice teachers. This observation about the clear difference in goals between the authentic and the simulated experiences was inspired by a similar observation made by Hopwood (2017) in his examination of simulation in the context of healthcare education. Hopwood argued that simulation does not have to seek to replicate everything about practice in order to achieve pedagogical outcomes. Indeed, the outcomes that the research is interested in are not those that would be set in an authentic marking operation; rather, they are pedagogic (or andragogic) goals of developing the volunteer participants’ knowledge, ability, competency, assessment literacy and/or confidence in making judgments.

From the theoretical perspective of PA, sayings, doings, and relatings, ‘hang together’ within the context of a project. ‘Hanging together’ is a term borrowed from Schatzki (1996, cited in Mahon et al., 2017); it is used in PA to refer to the ways in which several types of arrangements exist in a particular moment or situation (Kemmis et al., 2014), which is referred to as a site. The cultural-discursive arrangements are evident in the language used in the interactions between participants in the project; these are referred to in PA as sayings. The physical aspects of the project, the doings, are the material-economic arrangements of the project, and the relatings are the ways in which participant interactions are informed by the social-political arrangements between the people involved in the practice. The particular combination of arrangements present in an instance of a project are conceptualised as the ‘architecture’ of a process, hence the name of the theory. Further, the architecture of a particular process is also shaped by practice traditions – the ways in which the project has been done in the past can have a marked effect on the way participants carry out a project in a new iteration of the process. Kemmis et al. (2014) provided a diagrammatic representation of the architecture of a process, displayed in Figure 1.

In Chapter 8 of this thesis, which draws the various threads of the findings and results together into an integrated whole, the marking simulation as a project is considered with regards to the sayings, doings and relatings of the process. This will help in understanding how the marking simulation was able to achieve the outcomes that the researcher and participants observed and perceived.
Setting and Context

The research was situated in the state of Western Australia, where teachers are registered by the Teacher Registration Board of Western Australia (TRBWA). The TRBWA have adopted a modified version of the AITSL Australian Professional Standards for Teachers called the ‘Professional Standards for Teachers in Western Australia’ (Teacher Registration Board of Western Australia (TRBWA), n.d.). For teachers to achieve full registration, they must demonstrate that they meet the Professional Standards at the Proficient level (Teacher Registration Board of Western Australia (TRBWA), 2018). As in the AITSL Professional Standards, within Standard 5 of the TRBWA Professional Standards there are two areas that particularly focus on teachers’ assessment ability: Standard 5.1, ‘Assess student learning’ and Standard 5.3, ‘Make consistent and comparable judgments.’

Preservice teachers are likely to be aware of the need to refine their skills in assessing student work in order to achieve full registration. If not, then they will discover that they need to meet the TRBWA Professional Standards relatively quickly, as transitioning from provisional registration to full registration is functionally compulsory: it is expected within three years of graduation and the category of registration granted to graduates cannot be renewed except under “exceptional circumstances” (Teacher Registration Board of Western Australia (TRBWA), 2018).

Figure 1 The media and spaces in which sayings, doings, and relatings exist. From Kemmis et al. (2014, p. 34)
Design

The aim of the research lent itself to a longitudinal study, as its central concern was what change there may have been in the beliefs and perceptions of a group of individuals as a result of an intervention (the simulated marking experience). It was also concerned with whether these individuals’ perceptions or evaluations of any impact immediately after the intervention changed after they transitioned from preservice to beginning teachers. A multiple point prospective panel design (De Vaus, 2001) provided a structure for pre- and post-simulation data gathering that was able to capture not only immediate change in participants’ perceptions post-simulation, but also whether any perceptions of benefit remained after a period of time, once they had transitioned to employment. A single panel without replacement was deemed appropriate. It was not expected that all of the preservice teachers that were involved in the simulation would be able to secure employment as teachers immediately after graduation, and this introduced the potential that attrition would have reduced the panel to an unacceptably small sample. It was decided that, had that come to pass, the follow-up interviews would be developed into case studies (Punch, 2014; Silverman, 2010); however, this contingency plan was, fortunately, unnecessary. Almost a third of participants from the simulation were interviewed in the follow-up phase, which was considered to be a sufficient proportion.

In addition to the prospective longitudinal study, a cross-sectional survey of experienced teachers with large-scale marking experience was also undertaken following a retrospective panel design (De Vaus, 2001). The data collected from this retrospective study was intended to provide data about the beliefs teachers generally hold about what impact large-scale marking experiences have on teachers. These data are used for comparison with the post-simulation data in Chapter 8, with the hypothesis being that if the simulation is sufficiently accurate, the impacts reported by participants from the simulation will be similar to the impacts from authentic large-scale marking experienced teachers report in the survey. It must be noted that retrospective longitudinal studies are fraught with potential issues because of the inherent unreliability of human memory. Comparing the results of the retrospective study with reported impacts presented in the research literature, as well as taking the potential unreliability of self-reported beliefs into consideration in the data analysis, can address this concern.
The research was divided into three phases as shown in Figure 2 Structure of the research. The research began with the retrospective cross-sectional study, the ‘survey phase’ – phase one – which gauged whether the beliefs Australian school teachers have about the impacts of marking participation correlate with and/or extend the impacts reported in the literature review. This was intended to provide grounds to argue whether the various participants’ experiences of and beliefs about large-scale marking are likely to be generalisable across and beyond the Australian context. This phase was separate from the prospective longitudinal study, but it provided background and context for the longitudinal study that is not available from the literature, as well as findings in its own right. The second part of the research was the prospective longitudinal study, in two phases. First, the ‘simulation phase’ – phase two – where pre-simulation data was collected, the simulation was carried out, and then post-simulation data was collected. The final phase – phase three, also called the ‘follow-up phase’ – was collecting data from the participants approximately six months after they began teaching, seeking to understand whether they had reassessed the value of the marking simulation experience.

Figure 2 Structure of the research.

**Data Analysis**

Each of the chapters in this research that reports on a phase of the research contains the specific details of the data collection and data analysis conducted in that phase. However, in all phases, there is a common method of qualitative data analysis: the Interactive Model (Miles, Huberman, & Saldaña, 2014). So, it is pertinent here to explain this method as it will be referred to throughout the thesis.

Miles et al. (2014) described three ‘streams’ of data analysis activity, post-data collection: data display; data condensation; and drawing and verifying conclusions. Each of these streams of activity is separate; however, the researcher moves between them during analysis in order to refine each one. It is almost an iterative process, except that there is no
implied direction or order of working; one moves between each stream as necessary. The conceptual arrangement of the streams of activities are displayed in Figure 3.

![Diagram showing the interconnection of data collection, data display, data condensation, and conclusions: drawing/verifying streams.](image)


**Figure 3** The Interactive Model. Cited in Miles, Huberman & Saldaña (2014)

The data condensation stream involves cleaning data and adding attributes to cases and files, but the main activity in this stream is coding. There are multiple approaches to coding, and the Interactive Model is not prescriptive about which coding approach to use. In this research, coding was typically performed in two cycles. The first-cycle coding was preliminary coding, where the researcher identified themes raised in the data and selected a phrase that would concisely describe the general topic. The second-cycle coding then attempted to condense the first-cycle codes into fewer, more general codes.

Data display is a stream of activity where data is, obviously, displayed. Qualitative data analysis software – like NVivo, which was used in this research – is usually capable of many different visual displays, including networks showing relationships between cases and files, and word clouds, coding grids and inquiry reports. Sometimes, displaying the data highlights an aspect that encourages the researcher to further condense the data. At other times, it leads the researcher to the activity of drawing and verifying conclusions – particularly when they recognise a pattern in the data. In this research, one of the most common forms of data display was simply viewing all the references to a code in a single display, and read through to identify common aspects and attributes. Sometimes this led back to data condensation and a review of the codes; at other times, to drawing a conclusion about the theme/s present in the references.
SIMULATED LARGE-SCALE MARKING

The third stream of activity is to draw conclusions, and then verify them using the data. In the present research, this activity often coincided with drafting parts of the results sections of the middle chapters of this thesis. In writing summaries of the themes and sub-themes identified in the data, the researcher frequently returned to the data displays and the codes themselves to verify that the summaries accurately represented what was evident in the data. The results as they are reported are thus aligned with the coding and data displays the researcher used to understand the research participants’ experiences.

**Ethical Considerations**

An application for ethical approval was made to the Edith Cowan University Human Research Ethics Committee (ECU HREC) through the System for Tracking Research Ethics Applications and Monitoring (STREAM). This section gives a brief overview of the ethical considerations of this research in the following areas: informed consent; anonymity and confidentiality; and participant and researcher safety.

**Informed consent**

It is imperative that research participants are taking part voluntarily and that their decision about whether to volunteer is made based on a clear understanding of what will be required of them (Wiles, 2012). To this end, participants in each phase of the proposed research received a letter, prepared according to the guidelines provided by the ECU HREC, which clearly stated the aims of the research; the reasons for the research; how the participants’ data would be collected, stored and treated; what the participants would be required to do; the potential risks and benefits of their participation; and how anonymity and confidentiality would be handled. The letters also contained contact information for the researcher’s university supervisors and of an ECU ethics officer who was independent of the research. These letters appear at Appendix A.

Accompanying, but separate to the information letter, participants viewed (for the online survey) or received a consent form in which they could declare their understanding of the research, acknowledge that they may withdraw at any time without penalty, and affirm their consent to participate in the research. These consent forms were also prepared in accordance with guidelines provided by the ECU HREC, and they appear at Appendix B.

Participation in phase two of the research, the simulation, was considered to cause some inconvenience for participants, given that they were agreeing to take part in a whole-day activity. To compensate for some of the inconvenience of attending the simulation,
arrangements were made to provide morning tea and lunch at no cost to the participants. In considering the compensation, due diligence was paid to the ethical implications of how the compensation was communicated to participants and potentially perceived by participants and other parties. In particular, it was borne in mind that compensation should not be able to be construed as an inducement to participate, as inducements compromise the validity of participants’ informed consent (Wiles, 2012).

**Anonymity and confidentiality**

There were two different populations whose identity needed to be protected throughout this research, and whose disclosures must continue to be treated confidentially: the students whose writing is being used in the simulation, and the research participants. Each has different considerations, so each will be dealt with in turn.

The student writing, in the form ACER provided it, was anonymised as much as is possible with student writing. No names were systematically attached to the writing, only identification numbers. The identification numbers are linked to user identifiers in a confidential ACER database (which the researcher does not have access to), which are then linked to the students’ names in another database (also inaccessible to the researcher). In other words, the identification numbers that will be visible to the markers in the simulation are two degrees removed from the student’s identity.

However, it is possible that writing contains ideas or details that might identify or suggest parts of the student’s identity: the student may write in the first person using their real name, or may use a real-world setting that suggests their geographical location. During the simulation, participants brought to the researcher’s attention two different scripts where the student had included a name that was very likely to be theirs, as a signature or by-line. However, given the lack of additional information such as age or location, it is unlikely that even this amount of information was sufficient to precisely identify either student, especially given none of the schools that participated in the ACER research that collected the corpus of student work are situated in Western Australia. Nonetheless, the potential risk that students had disclosed information that needs to be kept confidential was identified prior to the simulation. So, the need to keep any disclosures in the student’s writing confidential was communicated to the participants in the information letter, explained in the marker training portion of the simulation, and reiterated when participants pointed out the scripts containing students’ names during the simulation.
In all three phases of data collection, participants were asked to disclose their thoughts and experiences. It is of the utmost importance that participants’ anonymity is assured by dissociating the data collected from their identity. This was done during data entry. All publications of results, including in this thesis, that reports on something a participant has disclosed, has omitted a label, used a pseudonym, or used a placeholder label (e.g. ‘Participant A’) to ensure that the participant’s name is never published with any of the collected data. Total confidentiality is not an assurance that a researcher can give to a participant, but anonymity is a necessary and achievable element of social research that can go some way to safeguarding confidentiality (Wiles, 2012).

The aims of this research were to collect data about professional experiences rather than personal experiences. This means that the data collected are less likely to contain sensitive or private information than research that aims to explore personal or private issues. This narrows the circumstances in which intentional breaches of confidentiality are likely to be necessary (Wiles, 2012). To mitigate the risk of unintentional breaches of confidentiality, all data has been stored and handled with the highest feasible level of digital security. Additionally, details in the data that have a reasonable potential to identify a participant (such as the name of a school where a participant has done a practicum) was altered before any publication of the data, including in this thesis, to help ensure both anonymity and confidentiality.

Safety

The risk of harm in all phases of the proposed research was quite low. There are two broad types of harms considered relevant to this research: occupational and emotional. Neither of these were considered likely to pose significant risks to the participants, but they were considered nonetheless.

During the simulation phase, there were some low-level occupational hazards associated with computer usage, but few other physical risks to the participants stemming from participation. To mitigate the potential occupational hazards of computer use, participants were reminded to monitor their own health and safety within the university environment (which they were all familiar with, given they were students enrolled at the campus where the simulation was held).

The risk of participants being exposed to potential emotional harm from their experiences in the proposed research was also considered to be very low. There were, in the simulation phase, potential impacts from stress associated with working under time limits,
feelings of uncertainty in making judgments, and perhaps feelings of lack of confidence. The researcher considered it his responsibility to be aware of these issues during the experience and take steps to mitigate them where necessary as they arose. This included adjusting the timing of breaks, and in two instances encouraging participants to take additional breaks when the participant felt that they needed them.

In contrast to the risks, it was anticipated that there would be benefits for the participants in the simulation: as has been noted, there appears to be limited opportunity in education courses for preservice teachers to read and score genuine student work. This research offered an opportunity for preservice teachers to read, interact with and discuss the quality of a range of student performances on a writing task. There is also the belief held by many teachers and researchers that large-scale marking experiences yields positive impacts. While it was too early to suggest that the simulation would have positive impacts similar to an authentic large-scale marking experience during the design of the research project, it was certainly the aim of this research to establish whether this may be the case. It was considered that there was at least a reasonable potential that the participants would gain some amount of professional benefit from the experience. These anticipated benefits were taken into consideration when weighing the potential risks and benefits to participants in phases two and three.

**Chapter Summary**

This chapter described the primary theoretical framework of this research – pragmatism – as well as a secondary theoretical framework, PA, that informs some of the discussion in later chapters. The three-phase design of the research was described, including the main activities in each of the phases. The method of data analysis that is used in all of the four subsequent chapters was also described. However, for clarity, specific methods used in each phase of this research are described in the relevant chapters: the survey design is described in Chapter 4, which reports on phase one; the quantitative methods used in analysing the scoring data collected from the simulation are described in Chapter 5; the interviews conducted in phase two are described in Chapter 6; and the follow-up interviews conducted in phase three are described in Chapter 7.
Chapter 4 – Phase One: Survey

As was pointed out in the literature review in Chapter 2, there has been little published research conducted in Australia on the potential or actual benefits of participation in large-scale marking, so it is yet to be established whether the findings from other places around the world hold true for Australian teachers. Of course, it stands to reason that similar activities in an Australian context would have similar impacts on teachers because the reported sources of benefit are conversations with colleagues about student work and exposure to a wide range of student work of varying quality (Gambell & Hunter, 2004; Goldberg & Roswell, 2000; Masters & Forster, 2000). That said, the extent to which these themes might be able to be generalised to the Australian context needed to be considered in the present research, and some indicative evidence gathered to argue that the literature related to other contexts holds true in the national and local context.

In order to do so, in the first phase of data collection Australian teachers who had participated in large-scale marking were surveyed by questionnaire. There is no necessary or fundamental connection between surveys and questionnaires (De Vaus, 2013). However, in this research, a questionnaire provided the most efficient means of data collection for the survey.

The primary aim of the questionnaire was to ask teachers who had previously participated in large-scale marking about their experience, particularly any benefits they believe came from it and any drawbacks. A secondary purpose was to collect perceptions of large-scale marking from teachers who had considered marking but not yet been able to do so. Taking both of these aims into account, three ‘paths’ through the questionnaire were designed, which would be determined by participant responses to key questions. This maximised the information gathered from all participants, as it considered the different experiences they might have had with large-scale marking. Further aspects of the survey design are discussed later in this chapter. The participants’ responses were then analysed to investigate whether it could be argued that Australian teachers have generally the same beliefs about the impacts of marking as their international colleagues, as reported in the literature.
SIMULATED LARGE-SCALE MARKING

Methods

This section explains the questionnaire design and participant recruitment methods for this phase of the research. It also provides specific details of the analysis methods for the data collected.

Survey Platform

There are best-practice guidelines in choosing a survey platform and designing survey instruments that broadly informed the selection and design of the questionnaire (cf. De Vaus, 2013; Sue & Ritter, 2012), though in keeping with the pragmatic nature of this research, ultimately the decision was made with attention to both theoretical and practical considerations. ECU makes Qualtrics available to its staff and research students. This survey platform is widely used in research and industry, as well as offering all the necessary functions and an acceptable level of security for respondents’ data.

Sampling

Consideration was given to sampling design, but it was decided that there was no feasible way to estimate beforehand the size of the population of teachers with marking experience, and so treating the ‘teacher-marker’ population as a discrete, closed population and attempting representative sampling was considered unlikely to be practical. Instead, a decision was made to try to gather responses from as many teachers with marking experience as possible.

Participant Recruitment

Initially, it was intended that the participants would all be recruited from Western Australia, as that is the local context for this research. However, during participant recruitment, a pragmatic decision was made to broaden recruitment to the national context in response to low numbers of responses and difficulties in the originally planned methods of recruitment. When this research project was proposed, it was intended to mainly recruit participants by approaching professional organisations for help, by forwarding a request to their members through their electronic mailing lists. Three professional associations were contacted, only one of which responded. That response indicated that the association would not forward a request by email to members, but that posts were welcome on their Facebook page. After a variation to recruitment methods was approved by the ECU HREC, three Facebook posts over the period from November 2017 to January 2018 were made on that
professional association’s page, as well as on another Facebook group page run by and for Western Australian English teachers. Qualtrics provides a link specific to social media so that responses collected by this method can be counted. In the period the survey was open, a total of 15 participants followed the Facebook link to the survey. No responses collected after February 1, 2018 are recorded as having come via Facebook.

In addition to the Facebook posts, some invitations to participate were disseminated with anonymous links through personal contacts and professional networks of the researcher and his main supervisor. The request was a brief introduction of the aims of the survey, with a letter of information attached, and a link to an online survey in the Qualtrics platform. These personal contacts contributed another five responses to the dataset.

The online survey went live in November 2017, and by March 2018 these social media and personal invitations had resulted in a total of only 20 responses. In order to gather a greater number, a decision was made to, firstly, expand the scope of the survey to include all Australian teachers, and also to publish a summary of the interim results though an online magazine with an accompanying link inviting participation (Reinertsen, 2018). The article was posted online on April 4, 2018, and links to it were disseminated through Teacher magazine’s weekly email summary. In April 2018, a total of 36 survey responses were recorded, none of which were via the social media link: all came through the anonymous link. It is likely, but not able to be proved, that almost all of these responses were as a result of the article.

From May to August 2018, a further seven responses were recorded. It is unclear as to whether these were further responses garnered through the Teacher article, or whether the link had been passed on by word of mouth.

The final sample size of the questionnaire was lower than anticipated. It had been hoped to gather enough responses to be able to support stronger arguments about the beliefs of Australian teachers about participation in large-scale marking. Nonetheless, the data that was collected was sufficient to posit likely answers to the relevant guiding research questions.

**Questionnaire Design**

For a questionnaire to engage the maximum number of participants, it must be relatively brief, engaging, and simple to access and answer (Sue & Ritter, 2012). These priorities formed the design principles for the questionnaire.

Due to the dual aims of the questionnaire, it was decided to filter respondents into two different streams based on whether or not they had participated in large-scale marking. One
SIMULATED LARGE-SCALE MARKING

of those streams was further screened, resulting in three possible ‘paths’ to the end of the survey. The underlying structure of the questionnaire is illustrated in Figure 4, while the items that comprise the questionnaire are found at Appendix C.

<table>
<thead>
<tr>
<th>Demographics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current or former teacher; Subject specialisation; Years of teaching experience</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Screening question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has marked</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Closed questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complex multiple-choice: impacts of marking</td>
</tr>
<tr>
<td>Simple yes/no: Have you considered marking?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Open questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reasons for marking</td>
</tr>
<tr>
<td>Expectations of marking</td>
</tr>
<tr>
<td>Advice to a potential marker</td>
</tr>
<tr>
<td>Barriers to entry</td>
</tr>
</tbody>
</table>

Figure 4 Questionnaire design.

The survey first gathered some background information about the respondent, namely how many years of teaching experience they had and their subject specialisation. The screening question was whether the respondent had marked in a large-scale assessment, and the respondent’s answer to this question determined whether they proceeded to the main survey or were directed to a different set of questions. From here on, these two groups of respondents are referred to as ‘markers’ and ‘non-markers.’

After the screening question, markers were directed to a question asking how many times they had marked, and whether they intended to mark again. These items were intended to gauge the markers’ experience with marking, and to perhaps begin to indicate whether they felt generally positive towards marking. From there, markers were presented with a checklist of impacts identified from the literature review and instructed to identify those that they agreed with. Two plausible distractors were included in the list, primarily intended to be a
means of gauging whether respondents were attentive and thoughtful in their responses. The implications of participants selecting these distractors will be discussed later.

The final part of the primary set of questions contained two retrospective questions and one general, reflective question in the form of a hypothetical. The two retrospective questions asked respondents to consider an earlier time in their career before they had any large-scale marking experience. It asked them to identify, in short, open answers, what reasons they remembered having for choosing to participate in large-scale marking, and what they had expected to experience. The final question asked markers what advice they would give to a colleague who was considering whether to mark.

The secondary set of questions aimed to gather perceptions of those teachers who had considered marking, but not yet marked – the non-markers. After the first screening question, a second screening question was presented to the non-markers, asking whether they had considered marking. Those who had considered marking were presented with an open-ended question asking them what had prevented them from marking, while those who had not considered it were presented with an open-ended question asking why they had not considered marking.

**Questionnaire data analysis**

The questionnaire collected data through both closed- and open-ended questions. The item-level data were pre-processed in Microsoft Excel. In the pre-processing, unneeded data, such as starting and finishing and time, and questions not seen by each participant, were removed. One incomplete response was also removed from the set of submissions; that response contained only one, single-word response to question 17. The data were split into two groups, ‘markers’ and ‘non-markers,’ according to whether the participants indicated they had previously participated in large-scale marking. Later in this chapter, the pseudonyms used for the responses reflect this division, using MR for ‘marker response’ and NMR for ‘non-marker response.’

Additionally, at this pre-processing stage, preliminary codes were generated for the responses to open-ended questions. These codes were recorded as headings in columns of an Excel spreadsheet. The cells of the rows of responses relevant to the codes were shaded with a colour, and a ‘1’ placed in them so that counts indicating the frequency with which each proposed code was present in responses could be automatically generated.

The NVivo qualitative analysis tool provides an ‘autocoding’ feature that can be used when importing data, and this feature was used after pre-coding, during the import of the
data. After verifying a successful import of the data, the automatically generated codes were examined for appropriateness and relevance to the research questions, and also compared to the set of preliminary codes identified in the pre-processing. The researcher observed that the automatically generated codes appeared to be encoding lexical frequency, which is a crude measure of the content of responses. As a result, the automatically generated codes were kept in the NVivo project, but ultimately disregarded when querying the data.

Subsequently, each response was read and coded. Firstly, the markers’ responses were coded, question by question, then the non-markers’ responses. The coding was done iteratively, in accordance with the Interactive Model (Miles et al., 2014). The preliminary codes were the starting point in the first-cycle coding, with additional codes being added in vivo. Then, a second cycle of coding was made to ensure that all responses had been compared to all codes. The final set of codes generated is displayed at Appendix D. There, it can be seen that the codes are categorised according to their ‘topic;’ that is, by the question the encoded responses were responding to. For example, the responses from markers were broadly categorised as ‘Reasons for marking,’ ‘Expectations of marking’ and ‘Advice for other teachers,’ in line with the three open-ended questions intending to gather information about these topics. This seemed a logical way to begin the coding, though it was observed that as the process developed, codes like ‘money’ and ‘time’ appeared in each of these three categories, indicating that there are broader themes that run across these ‘topic’ categories. This is discussed in more detail later.

Results

This section reports some of the general characteristics of the people who responded to the questionnaire. It also reports the responses to each of the open-ended questions in the questionnaire, and the responses to Question 11, a complex multiple-choice item. The results are presented with a small amount of interpretation, but a detailed comparison of the themes that are common across questions is left until the discussion section of this chapter.

General Characteristics of Respondents

Sixty-two complete responses were recorded: 43 markers and 19 non-markers. Additionally, as mentioned previously, one incomplete response was removed from the dataset. Of those respondents, 56 identified as secondary teachers: five former teachers, one new graduate, and 50 current teachers. Their teaching specialisations are displayed in Table 1. In addition to the secondary teachers, there were five primary teachers (four of them
current teachers) and one former early childhood teacher. The number of years’ teaching experience ranged from zero to 48, with a median and mode of 20, and a mean of 18.16 years with a standard deviation of 10.67.

Table 1 Teaching area specialisations of respondents.

<table>
<thead>
<tr>
<th>Secondary teaching area</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>26</td>
</tr>
<tr>
<td>No specialty identified</td>
<td>12</td>
</tr>
<tr>
<td>Humanities &amp; Social Science</td>
<td>8</td>
</tr>
<tr>
<td>Science</td>
<td>3</td>
</tr>
<tr>
<td>EALD</td>
<td>2</td>
</tr>
<tr>
<td>Digital Technologies</td>
<td>1</td>
</tr>
<tr>
<td>Maths</td>
<td>1</td>
</tr>
<tr>
<td>Music</td>
<td>1</td>
</tr>
<tr>
<td>Religious Education</td>
<td>1</td>
</tr>
<tr>
<td>Education Support</td>
<td>1</td>
</tr>
</tbody>
</table>

\(N=56\)

When separated according to whether the teachers had participated in large-scale marking or not, 3 of the non-markers were primary teachers, with the remaining 16 being secondary. There were two primary teachers and one early childhood teacher with marking experience, with the remaining 40 being secondary teachers. The non-markers had a range of 0 to 38 years’ teaching experience with a median of 10, and a mean of 12.47 with a standard deviation of 10.97. The markers had a range of two to 48 years’ teaching experience with a median of 20, and a mean of 20.67 with a standard deviation of 9.63. So, as a group, the non-markers were slightly less experienced in terms of the number of years they had been teaching, but there was a great deal of overlap, indicating that years of experience is not a good indicator for predicting whether or not a teacher has participated in large-scale marking for this sample of teachers.

The 43 respondents who had participated in large-scale marking before were asked how many times they had done so. Five respondents gave approximate answers, but these were taken at face value: for example, one respondent wrote “10–12” (MR43), and this was counted as 11; another responded “35+” (MR12), and this was counted as 35. If anything, reprocessing these five responses in this manner is likely to slightly underestimate of the number of times the markers had marked, and this was considered acceptable for the purposes of this research. One response was excluded from the following calculations: the respondent wrote “25 years” (MR27), which is too vague to be interpreted as there is no way of knowing how many times a year the teacher marked, nor to know whether the teacher did not mark in some years. The range of the number of times respondents had participated in
large-scale marking was 1–90 (and one non-response), with the median point falling between five and six. The mean was 10.79, with a standard deviation of 15.46. Excluding one outlier (a response of 90), the mean falls to 8.85 with a standard deviation of 9.18.

Respondents who had marked were also asked whether they intended to mark again. Only 42 responses were recorded to this question: 11 selected ‘no,’ four selected ‘maybe,’ and 27 selected ‘yes.’ The samples are too small to reliably compare against each other to quantitatively investigate whether there is a relationship between the number of times a marker has marked and their intention to do so again, but that would be an interesting question to investigate in a larger dataset.

On the whole, it would be fair to characterise the survey respondents as being mainly experienced secondary teachers who had participated in large-scale marking. This ought to be borne in mind when discussing the results of the survey, as non-markers, primary teachers and inexperienced teachers are underrepresented in the sample, meaning that their points of view may not be sufficiently present in the data to draw firm conclusions about the opinions and beliefs of these groups.

**Agreement with Statements**

All 43 respondents who had identified that they had participated in large-scale marking responded to a complex multiple-choice question that asked them to select which statements they agreed with from a set of nine statements about how large-scale marking affects teachers. Of the nine statements presented, seven had been derived from the literature and two were developed to be plausible distracters: statements that had not been identified in the literature, and that were not intended to be true. These statements are presented in Table 2, rearranged according to the number of times they were selected.

The analysis for this question was done in Microsoft Excel. A table was created with one response per row, and one statement per column. Where a marker had selected a statement, a ‘1’ was placed in the corresponding cell. This allowed automatic counts to be generated to summarise both the number of statements each respondent selected, as well as the total number of respondents who selected a statement.
Table 2 Statements about the effects of participating in large-scale marking.

<table>
<thead>
<tr>
<th>Marking a large-scale assessment …</th>
<th>Number of times selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>increased your confidence in your ability to judge the quality</td>
<td>40</td>
</tr>
<tr>
<td>of student work.</td>
<td></td>
</tr>
<tr>
<td>deepened your knowledge of the assessment itself.</td>
<td>40</td>
</tr>
<tr>
<td>supported your professional practice.</td>
<td>39</td>
</tr>
<tr>
<td>clarified the learning goals and expectations you have of your students.</td>
<td>38</td>
</tr>
<tr>
<td>raised your awareness of students’ literacy behaviours.</td>
<td>33</td>
</tr>
<tr>
<td>taught you more about students and their work.</td>
<td>29</td>
</tr>
<tr>
<td>deepened your discipline knowledge.</td>
<td>26</td>
</tr>
<tr>
<td>[distractor] developed your sense of professionalism.</td>
<td>17</td>
</tr>
<tr>
<td>[distractor] increased your repertoire of teaching strategies.</td>
<td>14</td>
</tr>
</tbody>
</table>

As can be seen in Table 2, no statement was selected by all respondents, and so there was no consensus among markers as to how markers are affected by participation in large-scale marking. However, all statements except the distractors were selected by more than half of the respondents, and there were three statements that were selected by 90% or more of the markers.

The three statements that 90% or more of markers agreed with were the statements about improving judgment, deepening knowledge about the assessment itself, and supporting professional practice. Given that marking is an exercise in professional judgment, perhaps the popularity of the statement about improving judgment is unsurprising. Also, as will be seen in the responses to the open-ended questions that follow, markers gaining insight into the assessment and its marking processes is a reasonably strong theme, and so the popularity of the statement about deepening knowledge is not unexpected. The high number of respondents that selected the statement about supporting professional practice is, perhaps, the most difficult to interpret, as the statement itself is vague. That it was so popular was surprising to the researcher despite it being derived from the literature, but on reflection and having considered the responses to the open-ended questions, it would appear that there are many perceived PD benefits of participating in marking. Perhaps, then, the statement about supporting professional practice tapped into the teachers’ beliefs about marking expanding and building upon teachers’ professional skills, and thus their practice.

Close in popularity to the three statements that were selected by 90% or more of markers, was the statement “clarified the learning goals and expectations you have of your students,” which 88% of markers selected. However, there is an apparent ‘gap’ between it and the next most popular statement, which sits at just under 77%. This may indicate an underlying structure to the pattern of responses that has not been identified in this research, or
SIMULATED LARGE-SCALE MARKING

it may be an artefact of the small sample size, and the apparent gap may not be significant. Further research with a larger number of respondents would be needed to draw conclusions.

The remaining three statements that had been derived from the literature were selected by more than half of respondents. This is a reasonably strong indication that what is true of the reports about the PD benefits of marking in other contexts might correspond very well with Australian teachers’ beliefs.

Of the 43 markers who responded to this question, nine indicated that they agreed with all of the statements, including the distractors. There were eleven respondents who selected one of the distractors in addition to their other selections. Each of the two distractors was selected by more than a quarter of respondents. There were four respondents who selected every statement except the distractors. There were no respondents who only selected distractors. Excluding the nine respondents who selected all statements, only one of the other 34 respondents indicated agreement with both distractors: all the others selected a maximum of one distractor.

In interpreting the results, it ought to be remembered that no question in the survey was compulsory, and so if respondents did not wish to answer, they could have simply ‘clicked through’ to the next question; they did not have to select any of the statements. Thus, the markers who selected all nine statements may have actually agreed with all, or they may have not given each statement the same amount of thought. It is also possible that the respondents who agreed with all statements might agree with any positive statement about large-scale marking, believing it to be beneficial in very many ways.

Distractors aside, there was a high proportion of respondents who indicated that they agreed with statements about large-scale marking increasing their confidence in making judgments, their knowledge of the assessment itself, and their professional practice in general. These three areas in particular should be kept in mind as the results from the open-ended questions are presented, because they provide a broad overview of the benefits from the marking experience reported by markers.

Reasons for Becoming a Marker

Question 2.5 of the questionnaire asked teachers who indicated that they had participated in large-scale marking, “Why did you decide to apply to be a marker?” This was one of the retrospective questionnaire items, asking markers to reflect back on the first time they marked. All 43 respondents who saw this question answered it, some with only a single word, and some with a few sentences.
SIMULATED LARGE-SCALE MARKING

The responses were coded according to the methods described earlier in this chapter, with all 43 responses first being coded as ‘Reasons for marking.’ These references were then coded again, with 10 more detailed codes – ‘sub-codes’ or ‘sub-themes’ – being created: see Table 3. Where responses presented multiple ideas, these were coded at more than one sub-code.

Table 3 Codes for ‘Reasons for marking.’

<table>
<thead>
<tr>
<th>Reasons for marking</th>
<th>Number of references</th>
</tr>
</thead>
<tbody>
<tr>
<td>Professional development</td>
<td>23</td>
</tr>
<tr>
<td>Gain an insider’s perspective</td>
<td>11</td>
</tr>
<tr>
<td>Money</td>
<td>11</td>
</tr>
<tr>
<td>Gain a wider perspective</td>
<td>8</td>
</tr>
<tr>
<td>Gain experience</td>
<td>3</td>
</tr>
<tr>
<td>Networking</td>
<td>3</td>
</tr>
<tr>
<td>Maintain links to the profession while being a primary carer</td>
<td>3</td>
</tr>
<tr>
<td>Minor themes (Invited; Recommendation from colleague; Interest; Job expectation)</td>
<td>1 each</td>
</tr>
</tbody>
</table>

Professional development. This was the code with the most coverage across the 43 responses from markers. This is an indication that PD is the most likely reason for teachers choosing to become a marker. Further evidence of this is found in the generally positive sentiment, that will be reported later. In order to better understand the various ideas and perspectives presented within this code, further passes of the data were made, resulting in five sub-codes being created within ‘Professional development.’ These are displayed in Table 4, along with their number of references.

Table 4 Sub-codes within ‘Professional development.’

<table>
<thead>
<tr>
<th>Professional development</th>
<th>Number of references</th>
</tr>
</thead>
<tbody>
<tr>
<td>Become a better teacher</td>
<td>7</td>
</tr>
<tr>
<td>Deepen understanding of their subject</td>
<td>5</td>
</tr>
<tr>
<td>Increase skill or competence</td>
<td>4</td>
</tr>
<tr>
<td>Deepen understanding of students</td>
<td>1</td>
</tr>
<tr>
<td>Increase confidence</td>
<td>1</td>
</tr>
</tbody>
</table>

In examining these sub-codes of ‘Professional development,’ one must keep in mind that whatever the specifics, each of these is an expression by a teacher that they believe there
SIMULATED LARGE-SCALE MARKING

to be some degree of PD benefit from participating in large-scale marking. This was hypothesised to be the case in the background to this study, based on the literature review as well as anecdotal evidence from personal communications and experience. In fact, some responses that were coded as referring to PD as being a reason for marking were not coded any further; for example, simple, clear statements like “PD” (MR25) or “professional development” (MR32). In these cases, of course, it is unclear what aspects of professional practice the respondents believed to be augmented by participation in marking, but they are clear and direct indications nonetheless of a belief that PD is a reason for marking. In fact, despite five sub-codes emerging in the analysis, all of the responses contain the idea that these markers chose to participate in marking because they were looking to improve one or more aspects of their teaching practice, or just generally become more effective teachers. The five sub-codes are described below, along with some illustrative extracts.

**Become a better teacher.** Perhaps the simplest or most straightforward of all expressions of PD benefit were from those teachers who made statements such as, “I believed the experience would benefit my role as a classroom teacher” (MR29), or, to “become a better teacher” (MR4). These statements lie at the core of why any teacher pursues PD opportunities. There were six references to this sub-code in the survey responses.

**Deepen understandings of their subject.** Five survey respondents made comments that were referenced to this code, including one of the science teachers, who wrote that they “wanted to understand more about the assessment and hierarchical nature of science skills” (MR10). This response is one of the few in the dataset to reference the hierarchical nature of skills that is the basis for making marker judgments. It is a small hint that one of the sources of professional benefit may lie in seeing the spectrum of quality produced by students.

When asked to consider large-scale marking, the respondents were given examples that included Year 12 exam marking. This background helps to understand some of the responses, particularly in this sub-code, where one of the respondents wrote, “learn how to be a better ATAR teacher by gaining a deeper understanding of the syllabus” (MR24). ATAR is the Australian Tertiary Admission Rank, a ranking for university entrance. It is also used to refer to the Year 12 courses that lead to students being given that ranking, hence the respondent describing themselves as an “ATAR teacher.”

**Increase skill or competence.** Four responses were coded to this sub-code because the respondent wrote about becoming more skilful or expert. One respondent wrote that they chose to participate in large-scale marking to “develop my expertise when marking my own students’ work” (MR23), while another expressed a desire to “improve professional
competence in assessment” (MR9). One expressly focused on specific skills, saying that they wanted to “become better at marking and improve my marking speed” (MR21). Another expressed a much broader desire: “I was looking for anything to help me develop some skills and knowledge” (MR41).

**Other codes.** There were three sub-codes within ‘Professional development’ that only have one reference each; however, the sample is sufficiently small that they still bear reporting. One of the respondents wrote that their reason for participating in marking was to deepen their understanding of their students, writing that they wanted to “learn more about the students so I could teach more explicitly” (MR30). Another hinted that validation of their judgment was a priority, saying they wanted the marking experience to “give me confidence in my own marking” (MR16). The third respondent was more general: “broaden my experience and knowledge” (MR22). As all of these three responses express an idea about developing professional skills, they still fall under the broader ‘Professional development’ code.

**Gain an insider’s perspective.** Eleven responses were coded to this node, all referring to gaining insight into the processes or content of the particular assessment being marked. This was characterised as an ‘insider’s’ perspective, because it seemed the respondents were expressing a desire to see how the assessment, or its marking operation, worked. One respondent felt that one way to understand the marking of the assessment was to participate in it, writing, “I’d tried to prepare students for the test in the past without really understanding how it’s marked” (MR7). This is an idea also expressed by MR6, who said they wanted to “see the exam from a marker’s perspective” (MR6). The reason, perhaps, for wanting the insider’s perspective is, as one respondent put it, to “improve my own students’ final scores and understand what the examiners wanted to read in the answers” (MR37).

In coding responses to this code, the researcher considered whether there was overlap between this specific knowledge of an assessment and the types of knowledge that fall under ‘Professional development.’ It was decided to keep these types of knowing separate, as the specifics of the marking process, rubric, and also what is valued by markers may not be generalisable to more generic professional competencies. On the other hand, wanting to see a specific assessment from an insider’s perspective is probably motivated by a desire to be a more effective teacher in the sense that doing so might enable one’s students to succeed on an assessment (which is expressed more or less explicitly in one of the already quoted responses). Perhaps the respondents, despite not expressing this idea in terms of PD, actually intended to express an idea about marking enhancing their professional practice. However,
because there is no clear way of knowing that, for present purposes these types of responses were considered different from each other.

Money. This code was clear and unambiguous: 11 direct references were made to money. Some responses were quite minimal, such as “money” (MR17). A few mentioned the idea in conjunction with other benefits: “Money but also the opportunity to further my skills” (MR2), “money and professional development” (MR32). Still others mentioned money after explaining other reasons, referencing it parenthetically: “(and the extra $$$)” (MR34).

It is a fact that large-scale marking is a paid activity, and it would be quite difficult to believe that if payment were removed as an incentive that people would be willing to spend weeks marking only for the other benefits they identify. Some of the respondents specifically refer to the income as supplemental to their other income, mentioning it explicitly as “extra money” (MR27), and none explicitly said that marking was their only source of income. The general indication from the respondents appears to be that the income is an added extra to other benefits, and not the sole reason for participating in large-scale marking.

Gain a wider perspective. Some of the differences between large-scale marking and the types of marking a teacher does within a school lie in the scope of the assessment. The variation in student ability within a class or school is likely to be much more constrained than the variation across a whole population, and so some teachers wrote that they took up marking opportunities to “gain first-hand knowledge of the standards across the state, and where my own students might sit” (MR1), or similarly, to “gain [a] better sense of teaching and learning across the state” (MR18). Another respondent, who also wrote about gaining an insider’s perspective, wrote of learning about the range of ability across the full spectrum of schooling:

I was also interested in reading work by students in year 3 and 5, as I’ve never taught primary school students. It was very interesting to see how their writing starts out. Since I mostly mark essays by students in years 11 and 12, it was useful to be reminded of just how far they come during their schooling! (MR7)

It is interesting to note that this marker was looking for both an insider’s perspective as well as to place his or her students in the broader context of ability, and it is important to reiterate that some responses contain multiple ideas about why the markers first participated in marking – none of the reasons for marking are mutually exclusive, and several respondents expressed multiple reasons.

This idea of seeing a range of ability in student work is relevant to a point raised earlier about one of the sub-codes within ‘Professional development,’ where a response was coded at ‘deepen understandings of their subject.’ In that response, the marker explicitly
SIMULATED LARGE-SCALE MARKING

mentioned “the hierarchical nature of science skills” (MR10). It is a slightly different idea to seeing a range of student ability, in that it is describing a conceptual progression within a domain, rather than a progression aligned with age or maturity, but the two ideas are broadly related in that they are about seeing a bigger picture of a learning progression. So, even though these ideas have been coded to different nodes in this analysis, the reasons teachers choose to mark are not necessarily independent of each other, or necessarily rigid in their boundaries. It is likely that among teachers who mark, there are interconnected layers of reasoning about the benefits they anticipate from marking opportunities.

**Gain experience.** There were three responses that referred to experience as being a reason for marking. One was a single word, but the other two both referred to being a new teacher. There was one who wanted to “learn from more experienced teachers/markers when I was a new English teacher starting out” (MR28) and one who stated, “I was a young teacher and wanted to broaden my experience and knowledge” (MR22). Both of these responses are reproduced here because the idea of participating in marking as being beneficial for beginning teachers is very relevant to the purpose and aims of this research. Obviously, two responses are not sufficient for drawing any conclusions, but it is worth noting here that the idea is present within the data that, firstly, marking has benefit, and secondly, those benefits may be available to new teachers.

**Networking.** Three responses referred to networking, two of them very briefly but one respondent writing, “to meet others teaching the course outside the institution in which I was working” (MR3). These few responses hint at the collegial nature of marking, and perhaps are a small suggestion about the community of practice that develops when working on moderation tasks with colleagues. This theme of collegiality was also identified in the literature review and conversations with colleagues as a possible source of some of the benefits teachers gain from marking (Masters & Forster, 2000).

**Maintain links to the profession while being a primary carer.** In contrast to the code above, which echoes the literature, three responses referred to something unexpected: using marking as an opportunity to maintain links to teaching when taking time out from the work force for maternity leave or to be a primary carer. The respondents wrote, “I was on maternity leave and going mad at home with the baby” (MR11), “I was a stay at home Mum and wanted to keep up to date with teaching without going into the classroom” (MR12), and “Was on maternity leave; wanted to maintain professional knowledge/development” (MR26). All three express a very similar idea, that marking is sufficiently close to the work of a classroom teacher that by participating in large-scale marking, one is able to keep a
connection to teaching when full participation in the work force is not possible. Again, three responses are too few for drawing conclusions, but the novelty of this finding is worthy of comment, and perhaps of further investigation in other research.

**Minor themes.** Four other codes were created in the coding, each with a single reference. It was decided to keep these codes as a way of identifying minor themes in the responses to this question. One respondent referred to being invited to mark, and another responded that it was recommended to them by their head of department. Another responded that exam marking was an expectation of their role at the time, and another wrote, “I enjoy the analysis, using different rubrics and making sure I am accurate so that the information being passed on to teachers is reliable” (MR15).

Again, it bears pointing out that there appear to be many reasons for teachers choosing to take up marking opportunities, but that none of those observed in this research are mutually exclusive with any others, and in fact several responses expressed multiple reasons. In other words, there is not consensus as to why a teacher chooses to mark, but prominent themes and recurring ideas emerge from this research.

**Expectations of Marking**

Teachers who indicated that they had participated in large-scale marking saw Question 2.6, which asked them, “Did you have any expectations of the marking experience?” This was the second of the retrospective questionnaire items, asking markers to reflect on the first time they marked. All 43 respondents who saw this question answered it. The responses were coded according to the methods described earlier in this chapter, and 11 codes nested within the *expectations of marking* node were created in order to group ideas and identify prominent themes. These codes and the number of references for each are displayed in Table 5. One code, ‘Negative emotional reactions,’ was created by grouping three codes from the first pass of the data together: nervous, intimidating and pressure.

<table>
<thead>
<tr>
<th>Expectations of marking</th>
<th>Number of references</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time consuming</td>
<td>11</td>
</tr>
<tr>
<td>No expectations</td>
<td>9</td>
</tr>
<tr>
<td>Negative emotional reactions</td>
<td>8</td>
</tr>
<tr>
<td>Beneficial</td>
<td>8</td>
</tr>
<tr>
<td>Difficult</td>
<td>7</td>
</tr>
<tr>
<td>Challenging</td>
<td>3</td>
</tr>
<tr>
<td>Tiresome or monotonous</td>
<td>3</td>
</tr>
<tr>
<td>Networking opportunity</td>
<td>2</td>
</tr>
<tr>
<td>Minor themes (Money; Insider’s perspective; Interesting)</td>
<td>1 each</td>
</tr>
</tbody>
</table>
**Time consuming.** The most common response to this question was that markers expected, before having marked for the first time, that participating in marking would be quite time consuming. Some respondents stated this simply, writing, “I knew it would be time consuming” (MR37), or that they thought it would be “demanding of time” (MR26). Others wrote about the difficulty of balancing the additional work against other commitments, writing “that it would be difficult to balance my marking/planning at work, with my external marking” (MR13), or that it would be “time consuming by the sheer volume you were expected to mark above and beyond your day job” (MR21).

The idea that marking takes a long time, or is time consuming, is not particularly surprising. Reading student work and making judgments is not a fast process. That a good portion of markers expected it to be so, and yet still chose to participate in marking speaks to this particular drawback being outweighed by those markers’ reasons for marking. The difficulty of balancing marking against other commitments is an idea that occurs in other parts of this research, particularly in the advice to potential markers, and so it will be revisited later, but for now it is worth pointing out that these respondents took this difficulty into account when deciding to become markers.

**No expectations.** The second most common response type was one of the simplest to recognise and code. These respondents had no expectations of what the marking experience would be before participating. There were eight single word responses: seven wrote “no,” and one more wrote “none.” Another wrote, “I had no expectations” (MR14).

Due to the responses being so brief, there is not much at this code to analyse, other than to point out that teachers taking up marking opportunities do not always know what the marking process will be like. This point may, in fact, be relevant to some of the other expectations that will be discussed below, where respondents wrote about the experience being intimidating or making them nervous.

**Negative emotional reactions.** Not all of the expectations of markers the first time they marked were optimistic. Some respondents reported being “nervous and unsure what to expect” (MR4), while others emphasised the stress involved in participating in large-scale marking for the first time using uncompromisingly negative language: “Stress/professional lack of confidence/dreaded the comparability session/stressed about moderation” (MR5). It is interesting to note that the respondent who wrote that comment (MR5) indicated in responses to other questions in the questionnaire that he or she had participated in large-scale marking 90 times and intended to do so again.
The source of negative emotional reactions about the prospect of marking appear to stem from the expectation that there would be comparisons between the markers and their peers. Alternatively, there may have been a fear of not meeting a standard; as one respondent put it, “I was very nervous that I would be ‘wrong’ or not good enough” (MR28). Such anxiety is, perhaps, understandable. Some of these negative emotions seem to result from what might, in other circumstances, be described as ‘peer pressure,’ or the pressure to conform. In evaluating these responses, one ought to bear in mind that these were reported expectations of the marking experience, not of the marking experience itself – it may be that these fears, nervousness and feelings of stress did not persist once marking had commenced.

**Beneficial.** Eight responses were coded as referring to benefits or value. All of the references were interpreted as referring to professional development benefits. Sometimes the reference to PD was explicit; for example, one marker wrote “That it would be very useful professionally” (MR16), and another thought that it would be “effective PD for me” (MR10). Another marker was more specific, anticipating that marking would help in “building collegiate understandings of my subject area” (MR3). In a few, though, the respondent was not explicit about the type of benefit, though it was inferred that these imply professional benefit: “That it would be time consuming but valuable (and it was both)” (MR2), and, similarly, “I expected it would be very time consuming but valuable” (MR38).

One marker wrote, “I only wanted to do it once, just for the experience” (MR6). That marker also responded that they had only marked once, and their advice to potential markers was to also “Do it once” (MR6). It is likely that this teacher still perceived the experience as having some benefit, given they recommended it to others.

It is perhaps unsurprising that the most commonly referenced reason for marking, that is, for PD, is also a theme in expectations for marking. In fact, the seven respondents who referred to benefit as an expectation had also given PD as one of their reasons for marking. These individuals in particular, perhaps, more strongly believe that marking is beneficial for markers.

**Challenging.** Three respondents referred to marking being challenging, two quite briefly, and one with slightly more explanation: “I knew that it would be challenging due to the moderation procedures, judging against exemplars and consistency” (MR31). The reason this code was not re-categorised under ‘Negative emotional responses’ was because none of these three responses appears to be negative about the challenge. The challenge being referred to is, perhaps, an intellectual one, though it is difficult to gauge sentiment in concise
responses like “challenging; interesting; demanding of time & expertise” (MR26), and one must be cautious about over interpreting such responses.

**Tiresome or monotonous.** This code also had three references in three responses, and they are difficult to interpret. Perhaps the most accessible one is in the context of a longer response, where the marker wrote, “I expected it to be a bit monotonous but I found that I enjoyed the routine and quickly picked up the pace as I gained more experience” (MR7), then went on to express other ideas about feeling nervous, which was coded differently. Another reference to this code was a simple, single word response: “tiresome” (MR9).

**Networking opportunity.** Two respondents indicated that they expected to “be able to connect with a range of other markers who could provide a wealth of knowledge” (MR36) and have “opportunities to network with other teachers” (MR22). This code is noteworthy in that it is another one that overlaps with a code in ‘Reasons for marking,’ highlighting again that there is a connection for some teachers between why they chose to mark and what they expected to get out of marking.

**Minor codes.** There were three codes with only a single reference each: interest, money and insider’s perspective. Importantly, all three of these are themes in the responses about reasons for marking, which is further corroboration that reasons for marking and expectations of marking are related to one another. In many ways, this may not be surprising because if one did not expect to gain something like “Money!” (MR13), then one would be unlikely to be motivated to volunteer and invest the time and effort in marking above and beyond one’s teaching workload.

**Advice to Potential Markers**

Question 2.7 asked teachers who indicated that they had participated in large-scale marking: “Finally, if another teacher asked you whether he or she should apply to be a marker, what advice would you give?” All 43 respondents who saw this question answered it, some with a couple of words, for example “Do it,” while others wrote a few sentences. This question is different to the ones preceding it. It was intended in the design of the questionnaire that the first questions were retrospective, reflecting on what the marker thought before marking, while this question was intended to gather responses that reflected the markers’ beliefs about marking at the time they did the survey. In this respect, the responses to this question ought to be evaluated differently because memory is not a confounding factor. It was not asking respondents to remember back any amount of time; it
was asking them for their current opinion of marking, framed as advice for a hypothetical colleague, and encouraged respondents to provide reasons why other people should mark.

The responses were coded according to the methods described earlier in this chapter, and were also coded according to sentiment. In addition, seven further codes were created within the ‘Advice to potential markers’ code; these and their number of references are displayed in Table 6.

Table 6 Codes for ‘Advice to potential markers.’

<table>
<thead>
<tr>
<th>Advice to potential markers</th>
<th>Number of references</th>
</tr>
</thead>
<tbody>
<tr>
<td>Professional development benefits</td>
<td>20</td>
</tr>
<tr>
<td>Warning about time</td>
<td>10</td>
</tr>
<tr>
<td>Networking</td>
<td>2</td>
</tr>
<tr>
<td>Gives insight into students</td>
<td>2</td>
</tr>
<tr>
<td>Minor codes (Minimum necessary experience; Pay not worth it; Wider perspectives)</td>
<td>1 each</td>
</tr>
</tbody>
</table>

**Sentiment.** During pre-processing, four sentiment codes were created: ‘Emphatically positive,’ ‘Affirmative,’ ‘Neutral,’ and ‘Negative.’ The first three codes had 14, 21, and 8 references respectively. The code for negative responses was deleted after the pre-processing stage as it had no references. This already gives an indication about the general sentiment expressed in the responses.

**Emphatically positive.** The 14 responses coded as *Emphatically positive* were those where the respondent strongly expressed positive encouragement through word choice and/or punctuation. There are short responses like “I would strongly recommend it” (MR37), and longer ones where only the first sentence of the response referred to this code, and the rest of the response was coded to another: “I think it should be part of every single teacher’s professional development, especially English teachers” (MR11). In one case, the response began with an undeniably emphatic statement, but went on to offer a caution or warning: “Yes, yes, yes! But be aware of the time commitment, it can be a crazy 2 weeks with little or no downtime, as well as making it difficult to stay on top of things at main job, so it may not suit those with young children etc” (MR22). The first part was still coded as *Emphatically positive*, while the warning or caution was coded separately at another node (in this specific case, to the code ‘Warning about time’). There were more of these mixed responses in the following code, so an explanation of how they were interpreted will be left to the next section.
**Affirmative.** This code was used to characterise 21 references where respondents expressed support for the hypothetical potential marker taking up a marking opportunity, but not in a way that was interpreted as emphatic. Many of the references were short responses; for example, “Yes” (MR41), “Do it” (MR17), and “Give it a go!” (MR23). Others were a little longer: “I would encourage them to participate, even if only for the professional development” (MR15).

Even if the response then went on to offer other ideas or even cautions, the first part was coded as affirmative. As an example, the first word of the following response was coded as an affirmative, while the rest of was coded as ‘Warning about time’: “Yes, if they're willing to give up 3-weeks of November to working evening and mornings” (MR18). Other responses were coded in a similar way, and Warnings about time was a commonly co-occurring code; for example, “Yes but be prepared to not have a life” (MR8), and “Yes – do it. But life is about marking and eating and sleeping” (MR27).

A reader might see a contradiction between an affirmative statement and a caution or warning. However, these contrasting ideas can be interpreted as an acknowledgment that participation in marking is worthwhile in spite of the drawbacks, costs, and negative parts of the experience. When interpreted in this way, it could be argued that even though the affirmative responses are less obviously positive, they are still expressing a strong sense of the inherent value of participating in marking, implying that these respondents see it as an experience that teachers should pursue, even if not all parts of it are good.

**Neutral.** There were eight references that were not explicitly affirmative, but neither were they advising against marking. Some responses were equivocal, such as, “Be prepared for three intense weeks. However, I feel that as pd it is invaluable” (MR1). Warnings about time were, again, a common occurrence; for example, “Only apply if you have extra time or not teaching anymore, otherwise it will negatively impact your teaching” (MR42).

The most negative sentiment expressed across all responses was MR31: “It depends on how much experience they have had. They will need to have been teaching for at least 5 years.” However, even this response does not discourage all teachers from participating in large-scale marking, and so was coded as Neutral. It is a response that raises some questions about why this marker thinks a minimum amount of teaching experience is necessary. Unfortunately, the idea was not found in other responses, and so the questions raised must go unanswered.
**Professional development benefits.** A major theme in the advice to potential markers is the PD benefits of participating in large-scale marking. The 20 references to it from the 43 respondents were further categorised into six sub-codes, displayed in Table 7.

<table>
<thead>
<tr>
<th>Professional development benefits</th>
<th>Number of references</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vague</td>
<td>9</td>
</tr>
<tr>
<td>Clarification</td>
<td>5</td>
</tr>
<tr>
<td>Improve own assessments</td>
<td>2</td>
</tr>
<tr>
<td>Minor themes (Increase marking speed; Increase confidence; Improve accuracy)</td>
<td>4</td>
</tr>
</tbody>
</table>

It can be seen that the sub-code ‘Vague’ has the most references; more explanation of the label is needed. The term was selected during analysis, but it has been since pointed out that it may not be as transparent as intended. ‘Vague,’ here, means that the respondents did not specify what they meant by PD in their responses, but did explicitly mention professional benefit or PD. In other words, the respondent mentioned PD while remaining silent about which aspects of professional practice were supported or improved. Two examples of this are, “I feel that as pd it is invaluable” (MR1) and “It is worthwhile professionally” (MR29). Even though the responses were coded as Vague, some quite clearly expressed a positive sentiment about marking: “it was some of the best PD that I have ever done” (MR39), or “these marking experiences have value added skills and are excellent PD opportunities” (MR10).

**Clarification.** Five responses were coded to this sub-code, and they are best characterised as referring to the marking experience clarifying curriculum, syllabus or subject expectations. Responses included “It allows you to understand exactly what the curriculum wants” (MR11) and “Gives better knowledge and depth of understanding of the syllabus - good professional learning” (MR24).

This sub-code and its responses have obvious similarities to the ‘Deepen understanding of their subject’ code in the responses about why the markers chose to become markers. Once again, it is important to evaluate these responses in light of being current, rather than retrospective, attitudes about and towards large-scale marking. That there is a similarity suggests that the retrospective responses might be coloured by current perceptions.

**Improve own assessments.** Two respondents wrote about how marking allowed them to improve their own assessments. One of them wrote, “Second major benefit is that large-scale marking helps to improve consistency in your own development and marking of school
based assessments” (MR14), and the other, “It has helped me see exactly what students need to do to achieve in there [sic] exams and tasks which I have then been allowed to pass on to my students” (MR40). These teachers were offering potential markers reasons to become markers, and it is interesting that none of the teachers gave this reason for becoming a marker in the first place. Improving teachers’ own assessments, then, may be an unexpected benefit of marking.

Minor themes. There were four responses that referred to PD in a specific way. One teacher noted that marking would help potential markers “develop more confidence in your marking ability” (MR35), and another asserted that marking “improves your ability to teach the skills needed, and it improves your ability to mark accurately” (MR13). One response referred to improving professional judgment, saying that marking “hones your ability to make judgments about student work instantly” (MR5).

All these minor codes characterise ideas that are unique among the relatively small number of responses gathered in this research, but that may be more widespread among teachers. It is important to note here the perspectives held by teachers about the PD benefits of marking are diverse.

Warning about time. There were 10 responses where markers included a caution about the time-consuming nature of marking, such as, “I would say yes, as long as you could commit to it properly - planned around it so you can make it a priority whilst involved” (MR16). As mentioned before, the warnings about time often collocated with positive statements: seven of the responses began with an affirmative, and one with an emphatic affirmative. In one case, there was an affirmative, a caution, and then a positive: “Do it, takes time but is enormously beneficial” (MR33). Perhaps the respondent wanted the advice to be overall quite positive, while acknowledging the time it takes as a serious drawback. The idea of balancing one’s commitments to work, family and marking appeared in seven of the ten references, with parts of the responses expressing ideas like, “I was a day marker and not juggling any other work or family commitments at the time. I think it would be very draining to do in the evenings after teaching during the day” (MR7) and “Yes but the work load is huge and it impacts greatly on yourself – tiredness, stress, interferes with family life” (MR34). The warning about time pressures causing stress occurs in one other response: “Know your reasons for applying to mark, as it can be a big time commitment and contribute to additional stress” (MR36).
SIMULATED LARGE-SCALE MARKING

The responses that contained a warning about time are clear in expressing it as a major drawback to being a marker. Their cautions are partly moderated by their clear encouragement to the hypothetical potential marker despite the drawback, but it would be wrong to assume that these markers thought that marking is a good idea for every teacher. They saw participation as being contingent upon having the right circumstances and mindset in order to balance the competing demands of ‘regular’ life.

**Networking.** There were two explicit references to professional networking in the advice to the hypothetical potential marker. These references were coded because networking appeared as a code within ‘Reasons for marking’ and ‘Expectations of marking.’ The fact that networking appears in all three categories of responses suggests that it is, perhaps, an important theme connected to large-scale marking. One of the respondents who mentioned networking in their advice to the potential marker wrote about marking “increasing networking skills” (MR30), and the other wrote about it being an “excellent way to work with other teachers” (MR26).

That networking again appears here may be partly or wholly explained by the observations made previously in this research that the marking panel is a Community of Practice (Wenger-Trayner & Wenger-Trayner, 2015), and that the source of many of the benefits of marking are conversations with colleagues about student work (Masters & Forster, 2000). It may be that these markers, and perhaps others, value the collaboration and discussions with peers highly.

**Gives insight into students.** There were two respondents who wrote about the marking experience helping them to understand their students. One of these references occurs at the beginning of a longer response, the rest of which was coded at ‘Improve own assessments’:

> It enables you to understand how students interpret the questions they are asked but also, through their responses, gives you an insight as to how they were taught. It has helped me to identify areas of misconception that students may have about a topic based on how I teach it and to alter my teaching to address this. (MR14)

The second reference is much shorter, but still expresses an idea about marking being a way to learn more about students: “it give you a valuable insight student [sic] behaviour in exams” (MR9).

One of the broad themes Falk and Ort (1998) identified was learning more about students and their work, and these few references in the advice that these markers would give to a prospective marker is a suggestion that this idea is present amongst Australian teachers as well.
**Minor themes.** There were four additional ideas that were assigned codes. One response was quoted previously as being the most negative: “It depends on how much experience they have had. They will need to have been teaching for at least 5 years” (MR31). This is the only response that indicated a belief that there was a threshold requirement for becoming a marker, and it was coded as ‘Minimum necessary experience.’ Another response was coded as ‘Pay not worth it’ and was the only one that indicated that the professional benefits from marking outweigh the financial remuneration, which in that respondent’s opinion was not sufficient: “The pay is not worth it but it is excellent PD” (MR38). The final minor code was created for a response that expressed the idea that marking allows teachers ‘Wider perspectives’: “VERY useful for moving out of the "bubble" teachers can get into, especially if they've been in a school for a long time/teach in an affluent school - creates awareness of/empathy for others” (MR26).

Once more, it appears that there are many reasons for, and expectations and experiences of marking. The uniqueness of the marking experience and its dependence on contextual factors such as the people one is working with and the assessment one is marking, has been noted previously in this literature review contained in this research (e.g. Goldberg, 2012), so these multiple ideas are not wholly unexpected. Even so, many of the perspectives identified in the markers’ advice to a potential marker occur in the Reasons for marking, the Expectations of marking, or the wider literature, supporting the idea that teachers understand the marking experience in similar ways, even though they may each focus on different aspects due to their personal preferences, values or opinions.

**Barriers to Marking**

Two questions (3.1 and 3.2) were asked of teachers who indicated that they had not participated in large-scale marking: to those who had considered participating, “What has prevented you from participating in large-scale marking?,” and to those who had not considered it, “Why have you never considered participating in large-scale marking?” The 19 respondents who saw one of these two questions all wrote a response, some writing up to a few sentences. The responses to the two questions were analysed together, as they all present ideas about what other than lack of willingness prevents teachers from marking. For the teachers who responded that they did not have the desire to mark, it is hypothesised that the barriers to marking outweighed the perceived potential benefits. For those who identified that they would like to mark but had not, the hypothesis was that those teachers could see benefit in large-scale marking, but still the barriers were insurmountable. It is important to note that
none of the respondents indicated that they had applied to be a marker and not been accepted. This underlines the agency of the non-markers in deciding whether it is ‘worth it’ to be a marker. The responses were coded according to the methods described earlier in this chapter, with six codes being created. These are displayed, with the number of references, in Table 8.

Table 8 Codes for ‘Barriers to marking.’

<table>
<thead>
<tr>
<th>Barriers to marking</th>
<th>Number of references</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>9</td>
</tr>
<tr>
<td>Distance</td>
<td>2</td>
</tr>
<tr>
<td>Lack of experience</td>
<td>2</td>
</tr>
<tr>
<td>Lack of opportunity</td>
<td>2</td>
</tr>
<tr>
<td>Lack of confidence</td>
<td>2</td>
</tr>
<tr>
<td>Other barriers</td>
<td>6</td>
</tr>
</tbody>
</table>

It is worth mentioning at this point that the overall purpose of the current research is to examine whether a simulated large-scale marking experience can offer similar benefits to authentic marking, and one of the justifications for simulating the experience is that simulations can be tailored in ways that authentic marking operations cannot. So, in examining the barriers to entry identified by survey respondents, ideas emerged about the ways in which a simulated marking experience might be tailored to accommodate the needs of teachers who ordinarily cannot participate in large-scale marking. These results are therefore quite important to achieving the overall aims of this research.

**Time.** It should not be surprising that time was a major theme in the responses to this question, as it has been in the responses to all other questions previously examined. References to time appeared in half of the 18 responses to this question. One of the ideas expressed was that it would be too difficult to be a marker and manage one’s other commitments. One respondent pointed to “time constraints mainly. Too difficult to mark large batches of work, work full time and manage a family” (NMR11). Another observed, “time constraints are the only thing getting in the way. I just can’t see myself doing my job to the standard I expect of myself and taking on extra responsibilities as well” (NMR12). One of the longer responses offers some insight into some of the emotional impacts of time management:

A sense of being overwhelmed by the amount of marking and practice-essay marking I was already doing as a teacher, and a general dislike for the drudgery of it (which, incidentally, explains most of why I left teaching). I enjoyed the ‘relief’ from this that the end of year 12 classes brought and was not a ‘glutton for punishment’, as they say. (NMR7)
SIMULATED LARGE-SCALE MARKING

By this point, the idea that marking is time consuming will be quite familiar to the reader, having been identified several times before in responses from markers. Thus, it is not very surprising that those who would like to participate in marking find this one of the most significant barriers to participation. In addition to the perspectives of those who would like to mark, one of the two respondents who had never considered marking also identified it as a barrier to marking writing: “Marking already consumes much of the time I have inside and outside of work hours” (NMR1).

Distance. Two respondents mentioned living in rural areas as being the barrier that prevented them from participating in large-scale marking. For these respondents, distance was closely tied to financial cost, with one writing that the barrier was “distance, time and money. I work in a rural school and all marking opportunities are in Perth. The school doesn’t pay travel costs so it would be expensive to attend” (NMR8), and the other, “I live in the country and previously markers have to be within a reasonable travelling distance of Perth to enable them to go into the city, mark and return home. The expense of staying overnight makes doing large-scale marking from the country prohibitive. If large-scale marking ever moves online, then it will be possible for me to do this” (NMR17).

This second response raises a possible solution, though it also raises questions about whether the online marking experience is able to help create a Community of Practice (Wenger-Trayner & Wenger-Trayner, 2015). This was examined by Adie (2012), among others, who found that there were professional benefits in online social moderation; this supports the respondent’s assertion that online marking might be able to extend marking opportunities to more teachers.

Lack of experience. One respondent wrote a single word response about what had so far prevented them from participating in large-scale marking: “experience” (NMR3). The other marker who referred to a lack of experience was much more specific, writing, “I have but because I haven’t taught a Stage 6 (Year 11/12) course before I haven’t applied. In NSW you need to have taught it for at least 1 year” (NMR2).

These responses contain an idea that is central to the current research: that preservice and early career teachers are often not qualified to apply for large-scale marking opportunities. In this respect, as mentioned before and discussed in much more detail later, simulation may offer a way to tailor an experience to these teachers.

Lack of opportunity. Two respondents identified that they had never had the opportunity to participate in large-scale marking. One wrote that “the opportunity has not arisen” (NMR18), and the other that they “just haven’t come across the opportunity at the
right time” (NMR14). These respondents had a similar problem to those who lived in rural areas, and those without enough experience: access to marking opportunities.

**Lack of confidence.** Both of the references to this idea came in conjunction with another barrier to participation. One respondent was very brief: “time and lack of confidence” (NMR10). The other wrote more, and did not explicitly refer to confidence, meaning the coding of the relevant part of the response relied on the researcher’s interpretation: “I also feared being slow and ponderous as a marker” (NMR7). The fear of not being able to mark at the required pace described here was interpreted as a reference to the respondent’s lack of confidence in their own abilities.

**Other barriers.** There were six references to other barriers to entry across four responses. One of the respondents who had considered marking referred to a conversation with another teacher being the factor preventing them from marking: “speaking to a colleague after they had completed Naplan marking” (NMR6). It is unfortunate that there was not more information about that conversation, as it raises a question about what that colleague reported about NAPLAN marking that dissuaded the respondent.

Another respondent mentioned time, then went on to write that they were “unsure about quality of feedback provided to markers” (NMR13). Again, it is unfortunate that there is not more information provided, as it is difficult to interpret this response. One possible interpretation is that receiving feedback on marking performance would be a primary reason for participating, and that if that feedback were not high quality, there would be less reason to participate.

Two of the responses contained multiple ideas that had not been coded to other nodes alongside ideas that had been. In fact, a portion of one of these responses has already been reproduced earlier in the appropriate sections (Time, and Lack of confidence). After commenting about the increased workload and their fear of not being able to keep pace with the required marking speed, the respondent in question wrote that they feared “potentially feeling conflicted by what the marking guidelines/rubrics/criteria appeared to value, and what I myself might have felt more important and worthy of rewarding in a student’s essay” (NMR7). This idea of there being a mismatch between the priorities identified in an assessment instrument and the marker’s own values in regard to what is important in a student’s writing is unique in this survey’s data.

The other response that contained multiple ideas was that of NMR9. This respondent also began with an acknowledgement of time being a barrier to marking, then elaborated:
Time. While I believe that the marking will give me a greater insight to student misconceptions quickly. I don’t believe that this is the only way to gain this knowledge. This can be done by analysing your student data effectively. I also believe it helps teachers learn how to teach to the test which is unfortunately what seems to be required in ATAR Mathematics. (NMR9)

Other than that time is a barrier, there are two ideas here: that there are other ways to gain insight into student learning, and that high-stakes assessment places pressure on teaching.

As already noted in reporting responses to the other open-ended questions, respondents offered many perspectives about the factors that prevented or deterred them from participating in marking. On the whole, though, there are some similarities in these responses to ideas present in the reasons for, expectations of, and advice about marking from those who have participated in large-scale marking. These similarities are discussed in the next section of this chapter.

**Discussion**

It will be remembered from the literature review in Chapter 2 that research from the US and New Zealand has described a range of potential benefits of participating in large-scale marking. The benefits for teachers reported in the literature include gaining awareness of literacy behaviours of their students (Gambell & Hunter, 2004), validating their expertise (Gilmore, 2002) and making them “more reflective, critical and deliberate” (Goldberg & Roswell, 2000). One article (Falk & Ort, 1998) identified four broad areas of professional learning that teachers derive from marking, though not necessarily large-scale marking: clarifying goals and expectations; deepening their discipline knowledge; learning more about students and their work; and developing insights that support their professional practice. The majority of the research done in this area reports benefits that mostly align with these four broad areas, though not all areas are reported in all studies. The variability of the reported benefits may be due to scoring experiences being unique in some respects, due to who the participants are, which assessment is being marked, or the characteristics and knowledge of the trainer. In other words, the variability in experience results from variability in the contextual factors of the marking experience (Goldberg, 2012).

The results from this survey of 43 Australian teachers were collected to help answer two of the guiding questions of this research: “What benefits do experienced Australian teachers report they have gained from participation in large-scale marking?” and “What are the opportunity costs, difficulties or other impacts reported by Australian teachers about their
participation in large-scale marking?” This section examines the results in light of these two questions.

With regards to the first guiding question, the four broad areas of professional learning from marking Falk and Ort (1998) identified provide a structure for examining the survey results. In this section, then, the themes that emerged in each of the sets of responses to the open-ended questions are reviewed in light of the four broader categories Falk and Ort identified: clarifying goals and expectations; deepening discipline knowledge; learning more about students and their work; and, developing insights that support professional practice. It must be kept in mind that these categories are not held to be definite or correct; there was at least one dissenting view identified in the literature. As noted above, Goldberg (2012) suggested that differing contextual factors for marking experiences results in differing impacts. This makes generalisations difficult, if not impossible.

Next, a synthesis of the perceptions with regards to the challenges and drawbacks of participating in marking is presented, in order to address the second of the guiding questions above (“What are the opportunity costs, difficulties or other impacts reported by Australian teachers about their participation in large-scale marking?”). In particular, this section reviews the expectations of, and barriers to marking that were reported in the survey, looking for the most common themes. The relevance of these barriers and drawbacks to this research is also briefly discussed, with particular reference to how simulation might mitigate some of these issues.

Ultimately, this section advances an argument about the extent to which Australian teachers’ perspectives on participation in large-scale marking reflect those reported elsewhere in the literature.

Clarifying goals and expectations

The complex multiple-choice question presented to markers contained two statements that are directly relevant to this area of professional learning: “clarified the learning goals and expectations you have of your students” and “deepened your knowledge of the assessment itself.” Thirty-eight and 40 teachers, respectively, agreed with these statements. In addition, the codes ‘Gain an insider’s perspective’ in ‘Reasons for marking’ and ‘Clarification’ in ‘Advice to potential markers’ both contain references to understanding the exam, curriculum or syllabus expectations. So, it is clear that this is a theme in the results.

This theme was reported by “an overwhelming majority” of Falk and Ort’s survey respondents (1998, p. 61), as well as in later interviews and observations. Gilmore (2002)
SIMULATED LARGE-SCALE MARKING

reported that her research in New Zealand also broadly aligned Falk and Ort’s (1998) findings. In regard to clarifying goals and expectations specifically, there are responses from teacher-markers in Gilmore’s research about the value of seeing the standard-setting process. Goldberg (2012) also noted that, in the specific context being investigated, teachers reported that marking “increased understanding of the standards and objectives” (p. 43).

The presence of this theme in the results of the present research as well as in the literature is the first indication that the beliefs of the teachers surveyed in this research are similar to those of their international colleagues.

Deepening discipline knowledge

In the complex multiple-choice question presented to markers, 26 respondents indicated agreement with “deepened discipline knowledge.” This is much less than the level of agreement with some of the other statements, but it still represents more than half of the respondents to the question. Additionally, within the ‘Professional development’ code in ‘Reasons for marking,’ there were five references to ‘Deepen understanding of their subject.’ So, this theme of learning more about one’s subject area or discipline is, perhaps, less strongly present in the data than some other themes. However, there may be unacknowledged overlap between this theme and some others. It ought to be pointed out that the categories used in this discussion may not be mutually exclusive, and so it may be that learning more about students, or constructing better assessments, or learning about the exam expectations for a subject should all be considered as increasing or deepening a teacher’s domain knowledge.

This theme was discussed only briefly by Falk and Ort (1998), and they gave no indication of how prevalent the theme was in the responses to their survey. It was not explicitly reported by Goldberg (2012), and also escaped specific comment from Gilmore (2002), although, as noted previously, Gilmore reported a broad alignment with Falk and Ort. In this light, perhaps it is less surprising that this theme was less prevalent in the present research than the other themes identified. In this way, the results of the survey reflect the literature.

Learning more about students and their work

Twenty-nine teachers agreed with the statement, “taught you more about students and their work,” and a further 33 agreed with “raised your awareness of students’ literacy behaviours” when presented with these statements in the complex multiple-choice question
near the beginning of the questionnaire. Additionally, the ‘Gain a wider perspective’ and ‘Deepen understanding of students’ sub-codes within ‘Professional development’ in ‘Reasons for marking’ combined for a total of another nine references to learning about students and their work. On the whole, it is a moderately strong theme in the data that teachers see marking as a way of learning more about their students.

It can also perhaps be argued that there is overlap between this theme of learning about students and that of clarifying goals and expectations, given that the exposure to a wide range of quality in student work is one of the ways in which expectations are set. Once again, it bears repeating that these categories are not universally accepted, and that they are not mutually exclusive; they are being used here to frame the discussion. The literature examining the benefits for teachers of marking in large-scale assessment programs all discusses and presents learning more about students and their work as a benefit of participating in large-scale marking (Falk & Ort, 1998; Gilmore, 2002; Goldberg, 2012). In that respect, the results of the present research align with the literature.

**Developing insights that support professional practice**

The strongest theme by far in all of the open-ended questions asked of teachers was PD, and in many ways it cuts across the three areas of learning discussed so far. PD was frequently referred to as a reason for marking, an expectation of marking, and also in the advice that markers had to offer to a prospective marker. Additional support for the theme was found in two statements from the complex multiple-choice question: 40 respondents selected “increased your confidence in your ability to judge the quality of student work” and 39 selected “supported your professional practice.”

This was by far the strongest theme to emerge in this research, but within it were many and varied responses about which parts of professional practice were enhanced or developed. The majority of third-level codes were created in order to understand the codes related to PD, simply because there were so many divisions and sub-groupings present within them. It seems likely that ‘professional development’ as a term is used for a very wide variety of learning activities, all of which may have varying levels of relevance and impacts, but all of which teachers believe ultimately benefit their professional practice. The findings of this survey correlate well with reports in the literature about how teachers have experienced large-scale marking in that a very large majority of markers say they believe marking is good for PD (Falk & Ort, 1998; Gambell & Hunter, 2004; Gilmore, 2002; Goldberg & Roswell, 1999–2000; Goldberg, 2012).
Minor themes

Several minor themes present in the responses gathered in this research are not easily classified under the four learning areas above. Though some of them are indicated in other literature, they were not reported by Falk and Ort (1998). A few of them are worth reiterating here, either because they are possibly novel and worthy of further research, or because they appear in only some of the literature reviewed thus far.

Maintain links to the profession while being a primary carer. First of all, the most surprising theme from the researcher’s standpoint was one within ‘Reasons for marking’: ‘Maintain links to the profession while being a primary carer.’ It was interesting to see three separate responses that all referred to using marking as a way to maintain links to teaching while being a primary carer or on maternity leave. This idea of ‘keeping up’ with the profession, or maintaining a skill is one that was not identified in the literature, yet it strongly resonates with the researcher’s personal experience, in that he sees his personal participation in marking continuing his link to teaching since he moved ‘out of the classroom.’ It is possible that the networking (discussed below) and collegial interactions that occur in large-scale marking are important for teachers’ identities and self-perception. It is this emergent theme, not identified in the literature, that the researcher believes is most worthy of future research. There may be a role for large-scale marking participation to either strengthen links to the profession or strengthen teacher identities in ways that could help to retain teachers in the profession.

Networking. This was another emergent minor theme that appeared in responses to all three of the open-ended questions presented to respondents. Whether the networking reported by respondents was professionally beneficial, personally beneficial, or both (as hinted at above, in terms of developing a teaching identity) is an open question. That said, this theme suggests strongly that teachers do form a Community of Practice (Wenger-Trayner & Wenger-Trayner, 2015) when working together as a marking panel. Collegiality and conversations with colleagues is a strong theme presented in Gilmore (2002), and is suggested there to be a primary source of the benefits associated with marking, citing Masters & Forster (2000).

Money. This is the last of the minor themes that will be discussed here, which is not to say it is the least of the minor themes present in the data. It appeared in responses to all of the open-ended questions; other minor themes appeared only in response to some of the questions. Whether in comments that the money was a reason to mark or in advice given to
prospective markers that the pay was not worth it (MR38, quoted previously), money was mentioned relatively often. Interestingly, money was also a barrier to marking for rural participants, in that it was expensive for them to travel to a city to participate in PD, and the cost for the length of time that large-scale marking takes was prohibitive, despite marking being paid. This last point is revisited in the next section, which discusses the barriers preventing teachers from taking up opportunities to mark.

Payment of markers is not something commented on in the literature, with only Goldberg (2012) referring briefly to markers in some large-scale assessments being paid to mark rather than it being part of their work. However, it should be borne in mind that the literature is largely concerned with what professional benefit may be gleaned from marking, not the reasons teacher chose to mark in the first place.

**Barriers and drawbacks**

The second of the guiding questions relevant to this part of the research is about the negative impacts reported by teachers. The barriers to entry reported by non-markers are also considered in conjunction with these negative impacts, as there are several important similarities. Importantly, few barriers and drawbacks are identified in the literature; again, it is more concerned with the impacts and effects of marking, not teachers’ incentives and reasons for marking, and so these findings are largely novel.

**Time.** The time-consuming nature of marking was a very strong theme in both markers’ and non-markers’ responses to the questions about expectations of marking, advice to potential markers, and barriers to marking. In particular, balancing the time needed against work and family commitments was a prominent concern. It is not wholly unsurprising that teachers who take on marking as ‘extra’ work, above and beyond their teaching loads, find it difficult to keep work and life balanced, just as it is not surprising that there are teachers who would like to be markers who consider it impossible because of their other commitments.

This theme has important implications for this research. Simulations are usually, if not always, condensed versions of authentic experiences. Whether it is microteaching, an online environment or a training session on a CPR dummy, many (if not all) simulations are shorter than the corresponding authentic practice. Thus, simulation has the potential to remove one of the most difficult parts of the large-scale marking experience, which should not be underestimated when evaluating the viability of simulated large-scale marking as PD. This evaluation is given in Chapter 9.
**Distance.** This minor theme is closely associated with the already discussed theme of ‘Money.’ There were three respondents who wrote that the barrier keeping them from taking up marking opportunities was their rural location, and they presented this barrier as involving both distance and money. There are (at least) two implications for this research and its focus on simulating marking experiences. Firstly, the simulation in phase two did not pay participants, and this was a key difference between the simulation and an authentic marking experience, as discussed in more length in Chapter 5. Secondly, simulations such as the one in this research are, essentially, ‘portable.’ There is nothing about the simulation that requires it to be in a specific location, unlike authentic marking sessions, which require careful selection of a site, satisfying a number of logistical requirements. The point that marking simulations may be a method for extending the benefits to rural teachers is revisited in the concluding chapters of this thesis.

**Negative emotions.** There were several codes created, mainly within ‘Expectations of marking,’ but also in ‘Barriers to marking,’ where respondents referred to stress, lack of confidence, difficulty, monotony, pressure or feelings of nervousness. These negative emotions were also occasionally referred to in ‘Warnings about time’ within ‘Advice to markers.’ The high-stakes nature of many large-scale assessments (e.g. final-year exams), as well as the duration of the marking operation compared to the high number of pieces of work to be marked does contribute to a stressful environment, and so the negative emotions reported are not surprising.

That markers reported these feelings about the experience yet overwhelmingly recommended the experience to others does, perhaps, alleviate some concerns about the strength of these emotions. However, it stands to reason that there will be teachers who would like to participate in large-scale marking who feel too intimidated to pursue an opportunity to do so. Here, then, is another implication for simulated marking experiences, where there may be no stakes associated with the assessment results, nor consequences for slow marking. The opportunity to participate in a simulation, without the pressure and associated negative emotions, may enable wider participation and build the confidence of teachers who would ordinarily shy away from the pressurised environment of a marking operation.
Chapter Summary

This chapter detailed the design and delivery of a questionnaire survey of Australian teachers, presented the results and discussed themes that emerged in the data. The survey informs the answers to the second and third research questions:

2. What benefits do experienced Australian teachers report they have gained from participation in large-scale marking?

3. What are the opportunity costs, difficulties or other impacts reported by Australian teachers about their participation in large-scale marking?

In connection to the first of these guiding questions, it was found that teachers’ responses to the survey revealed similar experiences of large-scale marking to those reported by teachers in the literature, and to PD benefits and experiences reported in other studies (Falk & Ort, 1998; Gambell & Hunter, 2004; Gilmore, 2002; Goldberg, 2012; Goldberg & Roswell, 2000; Palermo & Thomson, 2019). The very large majority of responses were positive and encouraging of others to take up marking opportunities, with the main reason being PD. The benefits broadly mirrored the four categories of benefit outlined by Falk and Ort (1998): clarifying goals and expectations; deepening discipline knowledge; learning more about students and their work; and developing insights that support professional practice. However, additional benefits were identified that went beyond those four categories, including maintaining links to the profession, networking and money.

In relation to the second guiding question, a range of barriers to entry for teachers who had not marked, as well as drawbacks reported by teachers who had marked, were identified in the survey results. The negative impacts of marking, that is, the drawbacks, were not found in the literature review. These included the time marking takes and balancing marking against other responsibilities, and negative emotions such as stress and anxiety. The barriers to entry identified included finding the time to participate and, for teachers outside metropolitan areas, distance.

These barriers and drawbacks are particularly relevant to this research, because a simulated large-scale marking experience might be designed to mitigate, or even completely remove, these downsides. This would allow opportunities for participation to be extended to teachers who in other circumstances, because of location, experience or cost, would be prevented from accessing the PD benefits so frequently cited in the literature and the results of this survey.
The main research question of this research is about the extent to which participation in a large-scale marking simulation can provide, for preservice teachers, a PD experience similar to authentic large-scale marking. The next chapter, Chapter 5, describes the marking simulation and presents the reported perceptions of the preservice teachers who took part. Their reports of the benefits and drawbacks of participating in the simulation are compared to the benefits and drawbacks identified in the international literature and in Australian teachers’ responses presented in this chapter, which have been shown to be broadly similar to the literature.
Chapter 5 – Phase Two: Simulation

The simulation of a large-scale marking experience for preservice teachers is the central focus of this research, and this chapter reports on the simulation that was run in August 2018 for a group of 22 preservice teachers. The aim of this phase of the research was to simulate a large-scale marking session to provide a learning experience that would contribute to the preservice teachers’ PD.

The methods, analyses and results presented in this chapter contribute to answering the fourth guiding research question: “Can large-scale marking be effectively simulated?” Of course, what is considered an effective simulation needs clarification. The traditionally prioritised goal of simulation is to faithfully reproduce a ‘real’ event, in order to give simulation participants an authentic experience from which they can learn. Hopwood (2017) problematised this ideal. He proposed that a high degree of fidelity is not necessary, perhaps not even desirable, in pedagogical experiences because the simulation itself is the site for practice and learning. The learning is not contained within some ‘real’ marking operation elsewhere that is being precisely recreated; the learning is at the site of the simulation itself and the practice that takes place there.

So, how should ‘effective simulation’ be defined in this research: is it a high-fidelity reproduction, or a hyperreal pedagogical experience? The dominant paradigm of this research, as discussed in Chapter 3, is pragmatism. This is in no small part because this research is rooted in a workplace practice that occurred and occurs as part of the researcher’s professional life. The marking guide, marking procedures, and student responses used in the simulation are not artefacts created for this research – they are genuine, and were used in a prior marking operation run by the researcher. Like these artefacts, the processes of the practice were also high-fidelity simulacra: in both the simulation and the authentic practice on which it was based, the researcher planned and facilitated the marking, supervising the markers as they scored. Nonetheless, there were differences between the authentic experience and the simulation, some explicit and anticipated (such as the simulation markers not being paid for their labour) and some unexpected. This will become apparent later in this chapter, and also in Chapter 6, which reports the participants’ perspectives on the simulation. In considering these differences, Hopwood’s perspective is pragmatically useful in that it offers a viewpoint from which deviations from ‘reality’ may not diminish the effectiveness of a simulation, thus allowing for more holistic evaluation of the extent to which large-scale
marking can be effectively simulated. This line of thought is revisited at the end of the chapter, in the discussion.

The first part of the chapter, Methods, explains the methods and design of the simulation in much more detail than the general overview provided in Chapter 3. In particular, more detailed demographic information about the participants, the marking design and marking methods are given here than previously. Additionally, the methods used by the researcher to collect data about his own experience of the simulation are reported.

The second part of the chapter reports on the analysis of data from two sources. The first source of data was the scores that were collected from the simulation. Some descriptive statistical analyses were carried out on these, as well as a Rasch analysis. The scores were not analysed to make any sort of argument about the adequacy of novice markers. Rather, the scores were examined in order to determine whether the training portion of the simulation was effective and authentic. The hypothesis being tested was that if marker training was carried out with sufficient fidelity, both marking operations ought to yield similar results, especially given the pieces of student work and the marking guide were common to both.

The second source of data analysed and reported was the researcher’s field notes, recorded voice memos and autobiographical recollections. As has been discussed previously, the idea for this research arose in the course of the researcher’s professional work, and he has professional experience in training markers and managing marking operations. Therefore, his perceptions of the simulation’s fidelity, while inherently subjective and open to bias, can offer a professional opinion about the degree to which a marking simulation is able to replicate a marking experience.

Methods

This section builds on the methodology and methods already presented in Chapter 3 by presenting in more specific detail the processes and procedures used in this phase of data collection. In particular, this section presents the design of the simulation, including the marker training and marking procedures.

Participants

Volunteer participants for the simulation were sought from among ECU education students in their final year who were studying a major or minor in a subject associated with literacy assessment: English, Humanities and Social Sciences, or Primary Teaching. The maximum number of participants was limited by practical considerations such as the
availability of computer resources and time, group dynamics and the need to have enough participants in the follow-up phase. Based largely on the availability of computer resources, that number was set at 26.

Potential volunteers were invited to participate through the researcher’s supervisors disseminating the information letter that appears at Appendix B. Consent forms were collected from all participants prior to any of their data being collected and these are held on file. There were some enquiries received from potential participants after the maximum number of participants had been reached. All respondents beyond the first 26 were advised that the simulation had met its maximum capacity. However, the first two students who could not be accommodated in the simulation were asked for contact details so they could be contacted in the case of a late withdrawal.

As it so eventuated, there were five students who gave their consent and from whom data were collected prior to the simulation (two in interviews and three via email), who withdrew before the simulation. The data they provided was permanently deleted and was not included in any of the analysis presented later in this chapter. The two potential participants who had provided contact details in case of such an eventuality were contacted, and one of them was available to participate in the simulation. So, there was a total of 22 participants in the marking simulation.

Of these 22 participants, 13 were completing a Graduate Diploma of Education (DipEd). The DipEd. is a one-year, graduate-entry qualification that is being phased out in Western Australia and replaced with a two-year qualification. The other nine participants were enrolled in a four-year Bachelor of Education (BEd) degree. DipEd students studying Secondary Education can specialise in two major teaching areas, or study a major and a minor specialisation. BEd students study a major and a minor. There was one DipEd student who was studying Primary Education, in which there is no discipline specialisation; the rest were studying to be secondary teachers. As a result of some students having two majors, the counts of teaching specialisations, displayed in Table 9, is much higher than the number of participants.
researcher’s Role. The researcher took a participatory role in the simulation. This decision was made because the researcher was well qualified to train markers and lead a marking session given his professional experience. The researcher was, at the time of the simulation, a Research Fellow at the Australian Council for Educational Research (ACER). In that role, the researcher gained experienced in managing all aspects of large-scale marking operations including marking design; marker recruitment; marking guide development, elaboration and exemplification; marker training; discrepancy reconciliation; and marking quality assurance. At the time of the simulation, the researcher already had experience in training and supervising the writing marking operation for a high-stakes Western Australian assessment, a low-stakes ACER-owned assessment, and had also trained markers internationally for a Southeast Asian regional assessment program.

The researcher was responsible for the logistics of the original marking operation upon which this simulation was based. That marking operation was run in January 2016, from ACER’s Perth office, and the researcher was solely responsible for recruiting, training and supervising the casual markers employed by ACER for that marking operation. He was also responsible for selecting exemplars and compiling them into the marking guide, which was adapted and used in the marking simulation.

Data Collection

Participants marked via an online interface that automatically saved assigned scores into a database, which was exported after the marking simulation. The scores provide some quantitative evidence of how effective the marker training was and of the participants’ marking ability. Each score that a marker assigns is representative of the marker’s cognitive processes leading to that particular judgment. While examining the marking performance of preservice teachers is not an aim of, or a question in, this research, the reliability of the participants compared to the pre-existing marking data may reflect the degree to which the simulation of marker training replicated the authentic marker training.
Time for taking field notes during the simulation was quite limited, because of the nature of presenting and training. However, once the participants had begun marking in earnest, the researcher was able to take a few minutes to reflect and write some brief field notes. These were done in a Livescribe journal, and so were digitised as they were written. The day of simulation was intense, so the researcher did not immediately reflect on the process, choosing to gather his thoughts first. So, the morning after the simulation, the researcher recorded two voice memos with his reflections on the simulation. The recordings themselves are not very long: one is six minutes long and the other nine, and so full transcriptions were not deemed necessary. The files were annotated in NVivo so that relevant comments and ideas could be readily accessed and only those portions selected to be reported in this research were transcribed.

**Time and Location**

The simulation was held in August 2018, at ECU’s Mt Lawley campus. The duration of the simulation was 9am to 4pm, including two 15-minute tea breaks (morning and afternoon) and a half-hour lunch break. This is comparable to a day’s marking in authentic marking operations, based on the researcher’s professional experience. Initial marker training took place from 9am to 10.30am, followed by a fifteen-minute morning tea break. There was further training from 10.45am to approximately 11.30am, then marking until lunch at 12.30pm. Marking continued from 1pm until 4pm, interrupted with a 15-minute break at 2.15pm, and a ‘refresher’ moderation meeting afterwards at 2.30pm. In total, most participants marked for approximately three-and-a-half hours. During the last hour, at approximately 3.30pm, six volunteers participated in a small-group interview while the others continued marking.

**Student Writing**

Authentic student writing is an essential element for a realistic marking simulation. Prior to the research commencing, explicit permission was sought from ACER’s CEO, Professor Geoff Masters, to use of a set of student writing collected as part of its eWrite assessment for the purposes of the research. The terms and conditions of the eWrite assessment contain a clause specifying that the data collected can be used and shared with third parties for research purposes. The terms and conditions, along with the email from ACER’s CEO giving permission for the use of the data were included in the application for ethics approval made to the ECU HREC, further details of which were presented in Chapter
3. Information about the student, including name and age, as well as the student’s school, were removed from the data prior to their use in this research: only a number identified each script, and no key for re-identification of the student writing was made available for this research.

**Marking Procedures**

The marker training was delivered according to practices similar to those outlined by Alderson, Clapham and Wall (1995) and Weigle (2002). These practices are common to many large-scale marking operations, and the researcher is familiar with them from his professional experience. In the original ACER marking session that is the basis for the simulation, a range of scripts was selected to exemplify the marking rubric, and they were compiled into a marking guide. That marking guide was reviewed and adapted for the simulation. The elaborated and exemplified marking guide was both an initial training document as well as a reference for markers during the marking session in both the original and simulated marking sessions.

The marking was carried out on-screen partly because many large-scale marking operations are now marked using this method, and so it adds authenticity. Additionally, the scripts currently exist only in digital form because the scripts come from an online assessment, and the original marking session was also conducted on-screen. The marking software used in large-scale marking is almost invariably proprietary. The proprietary marking platform used by ACER for the original marking was not available for the simulation due to budgetary restrictions. As a result, a simulated marking environment was developed using a database application, FileMaker 15, which allowed for scores to be recorded directly into a database. Consistent with authentic large-scale marking, only scores were recorded, not marker comments on the scripts. Screen shots of the marking interface are at Appendix E.

**Marking Design**

The marking simulation was designed with a rotational double-marking design, where each script is marked by at least two different markers. The design was to order the 537 scripts by reference number, divide the scripts into ten 50-script marking bundles (leaving 37 scripts unbundled), and assign one bundle per marker sequentially. So, if there had been 20 markers, each bundle would have been assigned to two markers: Bundle 1 would have been assigned to Marker 1 and Marker 11, Bundle 2 to Markers 2 and 12, and so on.
The size of the marking bundles was selected pragmatically, based on professional experience. The average marking rate of the scripts in the ACER marking operation was approximately 12 scripts per hour. On this basis, 50 scripts were allocated to each bundle. A proficient marker could mark this number of scripts in four hours but given that the simulation participants were novice markers, it was considered unlikely that any participant would complete a bundle in the marking session. This decision was made with the intention that every participant should have a more than adequate number of scripts to mark during the allotted marking time, thus would be able to participate for the full time of the simulation. The results, presented later in this chapter, show that no participants completed all 50 scripts, though one participant marked 41 scripts, which is a much faster rate than was anticipated in the design. This suggests that an allocation of 50 scripts was an appropriate quantity to ensure all participants had an adequate number of scripts to mark for the time allotted.

Unfortunately, the marking design was not able to be implemented exactly as planned. The first compromise of the design was a result of human error in dividing the sample into bundles: Bundle 1 comprised only 49 scripts rather than 50 (and so there were 38 scripts unbundled). The second compromise of the design stemmed from pre-assigning marking bundles to usernames, so that when two participants withdrew at short notice due to unforeseen circumstances, not all bundles were marked by two different markers. The last compromise in the design was the result of an unknown technical error in the marking system, where four participants were unable to log in. This was remedied after a short delay, but those participants were then assigned to different marking bundles than planned, which caused deviations from the planned marking design. Table 10 below displays the marker–bundle matrix that was the result of these compromises to the intended design, along with what percentage of scripts from the bundle were marked. Do note, however, that the marker numbers in this table do not reflect the log-in usernames; rather, they are labels assigned to the markers and used elsewhere in this research to identify the participants.
Table 10 Marking bundle assignments and percentage complete.

<table>
<thead>
<tr>
<th>Marker</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>61%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>02</td>
<td>66%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>03</td>
<td></td>
<td>24%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>04</td>
<td></td>
<td>48%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>05</td>
<td></td>
<td></td>
<td>48%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06</td>
<td></td>
<td></td>
<td></td>
<td>50%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>38%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>09</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>44%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>74%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>43%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td>72%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>82%</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>54%</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>32%</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>32%</td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>60%</td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>38%</td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>62%</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50%</td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>44%</td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>64%</td>
</tr>
</tbody>
</table>

As can be observed in Table 10, two marking bundles (3 and 8) were marked by only a single participant, while four other marking bundles (4, 5, 7, and 10) were marked by three participants. This is an unfortunate deviation from the original design, in which all bundles would have been marked by two participants. This has led to a situation where there is an unevenness in the data: some scripts were marked by three participants, while others were marked by only one. Needless to say, there were some lessons learned from the process; these are revisited later, along with their implications.

**Data Analysis**

The data analysis in this chapter can be divided into two parts: analysis of the quantitative scoring data, collected during the simulation, and analysis of the qualitative data, collected through the researcher’s field notes and voice memos.

The quantitative data collected were processed in Microsoft Excel for the analyses of measures of interrater agreement, and in ConQuest (Adams, Wu, & Wilson, 2015) for the
SIMULATED LARGE-SCALE MARKING

Rasch analysis that explored rater severity. Additionally, the original scores for the pieces of student writing were used as a comparison data set in the Rasch analysis to gauge rater severity against a benchmark. This is discussed in further detail later.

The qualitative data, collected through the researcher’s field notes and voice memos, were analysed according to the Interactive Model (Miles et al., 2014), which is used throughout this research and is explained in Chapter 3.

**Interrater Agreement**

Large-scale marking operations routinely analyse marks for interrater agreement and interpret this as an estimate of interrater reliability. This is to ensure, to the greatest extent possible, that the same piece of writing will receive similar marks from different raters (Brown et al., 2004).

As mentioned in the literature review in Chapter 2, there are three main types of metrics used to gauge interrater reliability: consensus, consistency, and measurement (Stemler, 2004). Consensus is often estimated through percentage exact and adjacent agreement; consistency is reported through correlational statistics; and measurement estimates are calculated through factor analysis methods or an Item-Response Theory approach such as the many-faceted Rasch model (Linacre, 1994). This research used percentage exact and adjacent agreement, Pearson correlations, and a generalised multifaceted partial credit Rasch model to explore interrater reliability.

Percentage agreement rates offer an insight into the extent to which an individual’s set of scores agrees with another set of scores, or in other words, to what degree there is consensus in the two sets of scores. In this research, the relevant percentage agreement calculations are between the scores of a simulation participant and the scores assigned in the authentic marking by a panel of trained, experienced raters provided by ACER. The set of scripts assigned to each simulation participant were not all marked by a single marker in the original marking operation, which had a panel of eight markers. Information about which markers marked which scripts was not available to the researcher for the purposes of this research. However, interrater reliability in the original marking operation was monitored and quality assured through reconciliation of discrepant marks, resulting in operationally acceptable levels of interrater reliability. It is not unreasonable, then, to consider the professional markers as interchangeable with each other – it should not matter which original marker scored which script because the marking panel had high levels of interrater reliability. Based on this assumption, the percentage agreement rates achieved by a simulation
participant can be taken to represent the degree to which that participant agrees with the collective judgment of the authentic marking panel.

In order to generate the percentage agreement rates, the markers’ scores were exported from the Filemaker marking system as Microsoft Excel spreadsheets. Each marker’s scores were kept on separate spreadsheets for the data analysis. Next to the markers’ scores, the existing scores for the relevant scripts were copied in, then the difference between the two sets of scores was calculated automatically by subtracting the existing score from the marker’s score using Excel’s formula function. An example of the calculation matrix from Marker_01 appears in Figure 5.

![Figure 5 Agreement calculations spreadsheet for Marker_01.](image)

As can be seen in Figure 5, calculations were performed in column AI, these being: in cell AI1 a count of the total number of cells in columns X to AF that contain a number; in cell AI2 a count of the number of cells in columns X to AF that contain a ‘0’ (indicating exact agreement); in cell AI3 a count of the number of cells in columns X to AF that contain a ‘1;’ and in cell AI4 a count of the number of cells in columns X to AF that contain a ‘-1’ (indicating adjacent agreement). From these counts, percentages were calculated, rounded to the nearest whole number, in cells AI6 and AI7, and those percentages summed, with the total being displayed in cell AI8.
A simple linear correlation of script total scores was used to initially evaluate the simulation participants’ consistency with the total scores for the same scripts produced by the ACER marking operation. The sample of scripts used for the purposes of this research (N=537) contained all possible total scores from zero to 28, though the same cannot be said for any single marking bundle that was assigned to a simulation participant. This is a limitation of the present research and could be considered in the marking design of any future marking simulations. As an example, Excel’s correlation function was used to calculate the correlation between the total scores in column L and the total scores in column W of Figure 5, with the correlation being displayed in cell AI10. This analysis was done for all markers, and the results are displayed in Table 11 in the next section of this chapter, which reports the results of all analyses.

Many participants expressed an interest throughout the course of the simulation about how their marking compared to the marks assigned to their scripts by professional markers in the original, authentic marking session. So, in the week following the simulation, the researcher computed the statistical measures of agreement using the method described above and emailed all participants using a form letter, shown at Appendix F, summarising their marking performance.

**Investigating Rater Severity with Rasch Analysis**

Consensus and consistency metrics are straightforward to calculate and simple to interpret: there are widely used heuristics to gauge quality (Brown et al., 2004), and at an even more basic level it is as simple as observing that bigger numbers are more desirable. However, there are more powerful forms of quantitative analysis that can be used to shed light on more complex aspects of marking. One group of methods, widely used in language assessment (including writing assessment) is Item-Response Theory and Rasch Measurement (Lumley & Brown, 2005). In this research, a Rasch analysis was performed in order to examine whether the preservice teachers were more or less severe in their application of the marking guide. A reliable marker is one who is able to apply the marking criteria consistently to all scripts, and so an investigation of severity is an investigation of marker reliability.

Georg Rasch (1960) formulated an elegant model based on the concept that the probability of a person succeeding on a task was a function of only the person’s ability and the task’s difficulty. This model has many useful properties, including that it can be used to construct objective measurement scales with invariant units of measurement, upon which both tasks and persons can be placed, allowing for objective comparisons to be made.
Rasch’s model, however, can only be applied to dichotomous data, while this research deals with the marking of writing, the scoring of which does not take the form of a series of ‘right or wrong’ judgments that produce 1s and 0s. In a writing assessment, the student only attempts a single task (responding to a writing prompt) and produces a single ‘answer’ to that task, and in scoring the response, criteria with varying numbers of score categories are applied. In other words, scoring is done by applying a number of individual scales to a single piece of work, and each criterion can be considered as a ‘task’ or item: one for spelling, one for ideas, and so on. Rasch’s model was extended into other models that can analyse polytomous data like that produced by writing assessments. One such model, a version of which is used in this research, is the Partial Credit Model developed by Masters (1982). The software used in this research, ConQuest (Adams et al., 2015), implements a ‘mixed coefficients multinomial logit’ model that is able to, with the correct constraints, generalise to a partial credit model (Macaskill & Adams, 2016). This partial credit model was one of the two models applied in this research to analyse the scoring data.

Linacre (1994) further extended Rasch partial credit models to be able to examine, among other things, the relative severity of different markers on a marking panel. His model is the many-faceted Rasch model mentioned earlier in this chapter – MFRM. In this model, test-takers, the tasks and the raters are conceptualised to be three ‘facets’ of the assessment that can be independently modelled. The ‘task’ facet can take the form of dichotomously scored questions, or rating scales or partial credit coding like those used in writing assessment. ConQuest can apply a multifaceted measurement model that is equivalent to Linacre’s, which Adams and Wu (2010) demonstrated with an earlier version of the software. The functionality has been retained in the version of the software used in this research.

The essential concept to understand about MFRM is that while Rasch (1960) modelled only person ability and item difficulty, Linacre (1994) devised a way to estimate as many more facets of the assessment as are deemed pertinent. In writing assessments, rater severity is often examined because there may be systematic differences in the strictness of one marker compared to another that may not be identified through correlational measures (Bond & Fox, 2007). If this were the case for any marker, the reliability of the marking panel is attenuated because a student’s score on their piece of writing would be different depending on the marker who scored it. Comparing the markers, then, by analysing the scores they gave, can be a powerful tool to explore rater reliability.
SIMULATED LARGE-SCALE MARKING

The relative severity of different markers is of interest in this research because one of the aims of the marker training portion of the simulation is to arrive at consensus understandings of the marking processes and marking guide. It is thus important to establish whether the simulated training was able to train participants to mark in a similar way to the professional markers who had previously marked the writing. Establishing that they were able to apply the criteria reliably speaks to the effectiveness of the marker training, which can be considered evidence of authenticity in the simulation.

Two analyses were run in ConQuest; the command syntax for these appears at Appendix G. The first analysis was a multifaceted model where one of the facets was ‘marking panel,’ to investigate whether the ACER panel and the simulation participants differed significantly when considered as groups. To model this facet, the scores were assigned a binary identifier: scores from ACER were assigned 0, and scores from the marking simulation were assigned 1. The severity estimates of the two panels were very close, and were within the SE of the estimates, indicating that there was no significant difference between them. Therefore, a second analysis was run without the panel parameter. This was, again, a multifaceted model but rather than considering the marking panels as groups, the simulation raters were treated as individuals, and the ACER marking panel was treated as a hypothetical individual rater. As noted earlier, this is a reasonable assumption as long as there is high interrater reliability.

Results

The results below are reported in three sections. Each offers different evidence for the ‘effectiveness’ of the marking simulation, which is the focus of the fourth guiding research question. The first section of results focuses on interrater reliability, and uses some simple descriptive statistics to examine the extent to which the simulation participants’ scores agreed with the scores previously assigned to the scripts in the authentic marking operation being simulated. This analysis provides information about the overall similarity of the results of the simulated and authentic marking operations, from which inferences can be drawn about the similarity of the conditions and, perhaps, the effectiveness of the marker training.

The next section focuses on rater severity, reporting the results of a MFRM analysis that includes a rater severity facet. This analysis provides additional information about the markers that can also be used to make inferences about the similarity of conditions and marker training, allowing for some triangulation of data when inferring the qualities of these
aspects of the simulation. Furthermore, it allows for a more empirical gauge of marking quality that may be of more general interest.

The final section is purely qualitative: the results of an analysis of the researcher’s field notes and of voice memos made shortly after the simulation. A small number of themes were identified, and the information was synthesised and summarised as further data about the ‘effectiveness’ of the simulation from the researcher’s perspective.

Quantitative Analysis of Interrater Reliability

Table 11 displays the number of scripts marked by each simulation participant, as well as the percentage of exact and adjacent agreement with the ACER marks for each marker. It also shows a correlation coefficient between the total scores for the scripts marked by the participant and the ACER marking panel. The mean number of scripts marked by simulation participants was 25.18 with a standard deviation (SD) of 7.96. It should be pointed out that four markers – 19, 20, 21 and 22 – were delayed in starting their marking by technical problems with the marking system, and this may have affected the average marking rate.

For 13 of the student markers, the percentage of exact agreement in scores is higher than the percentage of adjacent agreement. Across all markers, the lowest exact plus adjacent agreement rate was 78%, and the highest was 98%, with a mean of 91.55% (SD 4.91). The marker with the lowest exact plus adjacent percentage agreement rate (78%) was one of those delayed in marking, but who went on to mark 32 scripts; this may indicate the marker was rushing to the detriment of accuracy and suggests that this agreement rate may be an aberration.

Correlations of total scores of the scripts marked with their pre-assigned total scores were generally high, all being 0.72 or greater, with a mean of 0.84 and SD of 0.06. It is worthwhile reiterating that even though the rating scales for some criteria are very short (as few as three points), the correlation reported here is between total scores. It ought not to be the case, then, that the correlation is high only because of short marking scales per criterion, as it would be entirely possible for there to be disagreement in a number of criteria that would lead to different total scores.
SIMULATED LARGE-SCALE MARKING

Table 11 Marking summary statistics.

<table>
<thead>
<tr>
<th>Username</th>
<th>Number of scripts marked</th>
<th>Percentage exact agreement</th>
<th>Percentage adjacent agreement</th>
<th>Exact + adjacent agreement</th>
<th>Correlation with ACER marking panel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marker_01</td>
<td>30</td>
<td>53</td>
<td>41</td>
<td>94</td>
<td>0.81</td>
</tr>
<tr>
<td>Marker_02</td>
<td>33</td>
<td>53</td>
<td>41</td>
<td>94</td>
<td>0.80</td>
</tr>
<tr>
<td>Marker_03</td>
<td>12</td>
<td>42</td>
<td>49</td>
<td>91</td>
<td>0.89</td>
</tr>
<tr>
<td>Marker_04</td>
<td>24</td>
<td>44</td>
<td>49</td>
<td>93</td>
<td>0.81</td>
</tr>
<tr>
<td>Marker_05</td>
<td>24</td>
<td>34</td>
<td>49</td>
<td>83</td>
<td>0.72</td>
</tr>
<tr>
<td>Marker_06</td>
<td>25</td>
<td>45</td>
<td>47</td>
<td>92</td>
<td>0.75</td>
</tr>
<tr>
<td>Marker_07</td>
<td>19</td>
<td>51</td>
<td>38</td>
<td>89</td>
<td>0.83</td>
</tr>
<tr>
<td>Marker_08</td>
<td>12</td>
<td>56</td>
<td>38</td>
<td>94</td>
<td>0.78</td>
</tr>
<tr>
<td>Marker_09</td>
<td>22</td>
<td>58</td>
<td>37</td>
<td>95</td>
<td>0.87</td>
</tr>
<tr>
<td>Marker_10</td>
<td>37</td>
<td>53</td>
<td>40</td>
<td>93</td>
<td>0.90</td>
</tr>
<tr>
<td>Marker_11</td>
<td>21</td>
<td>55</td>
<td>39</td>
<td>94</td>
<td>0.81</td>
</tr>
<tr>
<td>Marker_12</td>
<td>36</td>
<td>46</td>
<td>45</td>
<td>91</td>
<td>0.77</td>
</tr>
<tr>
<td>Marker_13</td>
<td>41</td>
<td>55</td>
<td>40</td>
<td>95</td>
<td>0.79</td>
</tr>
<tr>
<td>Marker_14</td>
<td>27</td>
<td>40</td>
<td>42</td>
<td>82</td>
<td>0.85</td>
</tr>
<tr>
<td>Marker_15</td>
<td>16</td>
<td>44</td>
<td>50</td>
<td>94</td>
<td>0.85</td>
</tr>
<tr>
<td>Marker_16</td>
<td>16</td>
<td>44</td>
<td>49</td>
<td>93</td>
<td>0.82</td>
</tr>
<tr>
<td>Marker_17</td>
<td>30</td>
<td>56</td>
<td>40</td>
<td>96</td>
<td>0.9</td>
</tr>
<tr>
<td>Marker_18</td>
<td>19</td>
<td>61</td>
<td>37</td>
<td>98</td>
<td>0.91</td>
</tr>
<tr>
<td>Marker_19</td>
<td>31</td>
<td>54</td>
<td>42</td>
<td>96</td>
<td>0.94</td>
</tr>
<tr>
<td>Marker_20</td>
<td>25</td>
<td>51</td>
<td>39</td>
<td>90</td>
<td>0.90</td>
</tr>
<tr>
<td>Marker_21</td>
<td>22</td>
<td>42</td>
<td>47</td>
<td>89</td>
<td>0.81</td>
</tr>
<tr>
<td>Marker_22</td>
<td>32</td>
<td>37</td>
<td>41</td>
<td>78</td>
<td>0.88</td>
</tr>
</tbody>
</table>

The limitations of the small amount of data and rudimentary evaluation of them notwithstanding, the broad indication that can be drawn from the data is that they demonstrate acceptable levels of accuracy in the participants’ marking (Brown et al., 2004). This, in turn, speaks to the fidelity of the simulation of the marker training. It is evidence that the simulation participants were marking similarly to the markers who authentically scored the same pieces of work, which supports an argument that the marker training portion of the simulation was realistic. In particular, the data suggest that the training supported participants in developing shared understandings of the criteria in the marking guide and in accurately applying the marking rubric.
**Rasch Analysis of Rater Severity**

As described previously, two multifaceted Rasch analyses were conducted on the scoring data from the simulation combined with the original score from the ACER marking panel. In the first analysis, a ‘marking panel’ facet was modelled to identify whether there was a significant difference between the overall harshness of the ACER marking panel and the simulation marking panel. The severity estimates for the raters from the first analysis are displayed as a screenshot from the ConQuest output in Figure 6.

![Screenshot of ConQuest output](https://example.com)

**Figure 6 Analysis 1: Rater severity estimates.**

In Figure 6, there is a Standard Error (SE) displayed alongside the severity estimates for all raters, in the ‘Error’ column. This indicates the range within which the true location of the rater is likely to fall; the SE can thus be interpreted similarly to confidence intervals, as an amount ‘plus or minus’ to indicate the accuracy of the estimate. For example, the estimate for Marker 01 can be interpreted as -0.157 ‘plus or minus’ 5.183, meaning that the real location of Marker 01 could be anywhere from 5.026 to -5.34. This is an enormous range, and the SEs of all the severity estimates in this analysis were so large as to make any comparisons between individuals impossible. One cannot confidently infer any difference between raters.
SIMULATED LARGE-SCALE MARKING

based on this analysis. Similarly, the ‘marking panel’ facet was unable to confidently separate the panels from one another.

These findings indicate that there was no significant difference between the severity of the two marking panels: as groups, the simulation panel and the ACER marking panel appear to be similarly severe. Thus, a second analysis where the marking panel facet was omitted was needed.

A screenshot of the same display of estimates from the second analysis appears below as Figure 7. It is immediately obvious that the SEs of these estimates are considerably lower than those in Analysis 1. This indicates that the estimates are far more reliable, though there is still an appreciable margin of uncertainty in all of the estimates for the simulation participants. This is likely to be due to the small number of scripts marked by each marker. In comparison, the ‘synthetic’ ACER marker’s SE is very low because of the large number of scripts used to estimate the ‘perfect’ marker’s severity, which is very close to the 0 point of the scale.

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>UNWEIGHTED FIT</th>
<th>WEIGHTED FIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>rater</td>
<td>ESTIMATE</td>
<td>ERROR^</td>
</tr>
<tr>
<td>Marker_1</td>
<td>1.234</td>
<td>0.255</td>
</tr>
<tr>
<td>Marker_2</td>
<td>2.099</td>
<td>0.247</td>
</tr>
<tr>
<td>Marker_3</td>
<td>1.440</td>
<td>0.309</td>
</tr>
<tr>
<td>Marker_4</td>
<td>0.584</td>
<td>0.302</td>
</tr>
<tr>
<td>Marker_5</td>
<td>-0.650</td>
<td>0.282</td>
</tr>
<tr>
<td>Marker_6</td>
<td>-1.815</td>
<td>0.308</td>
</tr>
<tr>
<td>Marker_7</td>
<td>2.376</td>
<td>0.379</td>
</tr>
<tr>
<td>Marker_8</td>
<td>0.690</td>
<td>0.442</td>
</tr>
<tr>
<td>Marker_9</td>
<td>0.550</td>
<td>0.393</td>
</tr>
<tr>
<td>Marker_10</td>
<td>1.461</td>
<td>0.309</td>
</tr>
<tr>
<td>Marker_11</td>
<td>0.992</td>
<td>0.390</td>
</tr>
<tr>
<td>Marker_12</td>
<td>2.842</td>
<td>0.298</td>
</tr>
<tr>
<td>Marker_13</td>
<td>2.485</td>
<td>0.253</td>
</tr>
<tr>
<td>Marker_14</td>
<td>2.169</td>
<td>0.325</td>
</tr>
<tr>
<td>Marker_15</td>
<td>-0.948</td>
<td>0.392</td>
</tr>
<tr>
<td>Marker_16</td>
<td>-0.541</td>
<td>0.399</td>
</tr>
<tr>
<td>Marker_17</td>
<td>1.121</td>
<td>0.364</td>
</tr>
<tr>
<td>Marker_18</td>
<td>1.682</td>
<td>0.480</td>
</tr>
<tr>
<td>Marker_19</td>
<td>0.951</td>
<td>0.284</td>
</tr>
<tr>
<td>Marker_20</td>
<td>2.718</td>
<td>0.331</td>
</tr>
<tr>
<td>Marker_21</td>
<td>-0.554</td>
<td>0.511</td>
</tr>
<tr>
<td>Marker_22</td>
<td>2.636</td>
<td>0.388</td>
</tr>
<tr>
<td>Marker_23</td>
<td>0.162</td>
<td>0.088</td>
</tr>
</tbody>
</table>

An asterisk next to a parameter estimate indicates that it is constrained
Separation Reliability = 0.929
Chi-square test of parameter equality = 644.18, df = 23, Sig Level = 0.000
^ Empirical standard errors have been used

Figure 7 Analysis 2: rater severity estimates
SIMULATED LARGE-SCALE MARKING

It is sometimes difficult to obtain a clear picture of distributions from a table of data, and so visual inspection of the distribution is often far more informative and intuitively interpretable. A Wright map is frequently used in Rasch analyses to display the measurement scale, in units called logits, along with the locations of persons on that scale and any other relevant facets. A screenshot of the Wright map from the Analysis 2 output is displayed as Figure 8. The leftmost column displays the scale in logits, along with the distribution of students’ writing ability estimates, as derived from their writing performances. Higher numbers in this column represent more ability, and lower numbers represent less ability. In the middle column, the raters are displayed using their marker numbers, according to the estimate of their severity. Those appearing higher were more severe or harsh in their judgments, and those at lower locations were more lenient. It should be observed that although the locations in this column are interpreted as severity, the locations are on the same scale as student ability and item difficulty – the units of measurement are the same. The ACER marking panel is represented as marker 23 in that middle column. In the last column, the marking criteria are displayed at the location of the mean difficulty of their score categories, using labels referring to their order in the marking guide. These, again, are positioned on the same scale as the students and raters, with higher locations being more difficult criteria, and lower ones easier. It can be seen that the locations of criteria are aligned with the approximate centre of the distribution of students, and this is indicative of a well-targeted assessment.
As noted previously, Marker 23 is the ‘synthetic’ rater that comprises judgments from all markers on the ACER marking panel. As can be seen in Figure 8, Marker 23 is located at approximately 0 logits, which is the location of the mean item difficulty. In other words, the ACER marking panel as a whole was neither noticeably lenient nor severe. In comparison, it appears as if most participants were more severe than the ACER rating panel. These
locations, of course, need to be interpreted in light of the SEs for the estimates. Nonetheless, it is clear that seven of the student markers appear to be noticeably harsher than the ACER panel at between two and three logits, and another four are more than a logit harsher (Markers 1, 3, 4, and 18). One marker (Marker 6) appears substantially more lenient at approximately -2 logits. There are 10 markers, though, who are ‘clumped’ around the 0 location, +/- 1 logit. This indicates that these markers were moderate in their severity. With more data, which would increase confidence in all these estimates, we would be able to draw conclusions. At present, though, it can only be suggested that at least 10 out of the 22 judged with similar levels of harshness as the professional marking panel. For the purposes of the current guiding research questions, though, it is enough to be able to say that the reliability and severity of the simulation participants appear to have been fairly similar to the ACER marking panel, which implies the marker training in both conditions was similarly effective.

**Limitations of the Quantitative Analyses**

There are obvious limitations to the evaluation of reliability and interrater reliability described above, including the short period of marking time and the low number of scripts marked by each participant. This severely limits the ability to generalise about the marking accuracy of the participants, and in light of that, the results reported above should not be interpreted as an argument that novice markers are as accurate or reliable as experienced, professional markers. That hypothesis has received attention from other researchers (e.g. Attali, 2016; Cumming, 1990) and is tangential to the present research. Rather, the argument being put forth is that the simulated marker training was likely to have been as effective as the marker training in the authentic marking operation, given the similarity in reliability achieved across the two occasions. Again, given the small dataset, that argument is not particularly strong. However, there is a general indication that the training was effective in supporting reliability in the marking. This suggestion can be further supported by the qualitative perspectives offered by the researcher’s impressions, field notes and voice memos. These are the focus of the next sections of this chapter.

**Qualitative Analysis of the Researcher’s Perspective**

The researcher made field notes during the simulation, as time allowed, and also recorded two voice memos the day after the simulation in order to preserve his perceptions of, and reactions to the simulation experience before too much time had elapsed. It is the nature of self-reported data to be inherently biased, being as they are the opinions and
recollections of a single source. The usefulness of these perceptions, then, is not so much that they provide accuracy and objectivity, but that they provide an additional layer of understanding of the experience that may raise questions deserving of further investigation, or issues prompting discussion and closer examination.

As noted in the Methods section, the analysis of these sources was done following the Interactive Model of qualitative analysis. The Interactive Model is discussed in detail in Chapter 3, and has been used in all phases of this research. Due to the relatively small amount of data being analysed, identifying themes was slightly more difficult than it was in phase one of the research, for example, because there were few ideas referred to more than once within a source or referred to in more than one source. Ultimately, three themes were identified: Inter-marker conversation, Approaches to marking, and Logistics failures. Below, each of these themes is considered, then a summary of the researcher’s overall perception of the simulation is synthesised from them.

**Inter-marker conversation.** One of the aspects of the simulation condition that diverged most obviously from an authentic marking operation was the amount of discussion that occurred between participants during the marking portion of the day. The participants’ perspectives on the discussion is discussed in Chapter 6, but the researcher’s perspective is explored here. During the simulation, the researcher made the following note in his field journal:

![Figure 9 Extract from the researcher’s field notes.](image)

For clarity, the transcription is, “Participants are quite chatty, comparing and reading and discussing quality. Not authentic, but perhaps useful?” This note was made not very long after the simulation participants had begun the marking portion of the day, which is significant because it is evidence that the researcher identified the frequent discussion between markers quite early. Additionally, it is clear that the researcher was questioning relatively early in the marking portion of the simulation whether the discussion was likely to benefit participants.
In one of the voice memos made after the simulation, the researcher reflected that the decision not to discourage the conversations was a conscious one but that, “perhaps I should have spoken to it, at least, a couple of people have said it was very distracting for them. So perhaps I should have spoken up and said, ‘Guys, keep it to those decisions where you’re really uncertain about a judgment and need help.’” The researcher then went on to reflect on the difference between authentic marking and the simulation in terms of marker attitudes. The researcher noted that professional markers may have a different attitude to marking because they are paid employees, but also because they are aware that their judgments will be reported to students. So, there are consequences to their judgments that were deliberately absent from the simulation condition, where the marks and the marking were explicitly declared to be solely for the experience of the participants, with no consequences for the students who produced the responses being marked. The researcher also noted in the voice memo that he thought that conversation with peers is likely to be one of the key sources of PD for markers. He reflected on his own marking experiences in making this observation.

In the same voice memo, the researcher noted that in one of the post-simulation interviews, it became apparent that some of the participants in the simulation had social ties to one another, while also noting that professional markers are often working with people who they know only professionally and not socially. This recollection and observation is relevant for this theme because it shows the researcher was aware that the inter-marker discussion was a divergence from the authentic marking operation on which the simulation was based.

In short, the researcher’s perspective on the discussion that occurred in the marking room during marking was that it was unlike an authentic marking experience, but that because it might have been a source of professional learning or development, he decided not to discourage the discussions. This decision reduces the fidelity of the simulation, but as discussed later, whether that reduces the simulation’s effectiveness is open to debate.

Approaches to marking. In both voice memos, the researcher referred to marking approaches. In the first memo, it was in reference to the difference in attitude associated with the consequences of the scoring, in the context of reflecting on the inter-marker discussions that took place during the marking. In the other memo, the researcher spoke about the theoretical approach to marking that informed the marking rubric’s construction and application. There were two aspects of applying the marking rubric that he recalled having to discuss with participants at different times during the simulation. This first was an analogy he introduced in the morning marker training session in response to a participant’s question
SIMULATED LARGE-SCALE MARKING

about awarding half marks; the second was about independent judgments and the ‘halo
effect’ wherein markers grade all criteria highly because they are impressed in general by a
piece of writing, rather than making independent judgments for each criterion.

These references to marking approach can be interpreted as the researcher being
aware that the simulation participants brought different understandings and expectations to
the simulation than did the professional markers who were employed in the authentic
marking experience, and that he was aware not only of a difference in attitude and purpose,
but also a difference in operational understanding of applying a rubric. The first difference is
likely to be caused, wholly or in part, by the fact that the marking exercise was a simulation,
while the second difference is likely to be due to the participants themselves. The simulation
participants appeared to have very little experience with applying or using rubrics, despite
being broadly familiar with the idea of a rubric, and having been assessed themselves many
times with rubrics.

In summary, the inter-marker discussion and the attitudinal approach to marking were
differences between authentic and simulated marking that arose from the fact that preservice
teachers were participating in a simulation of marking. While the difference in rubric
application resulted from a difference between the participants themselves and the markers
employed for the authentic marking operation being simulated in this research.

**Logistics failures.** The third theme in the researcher’s notes and reflections was
related to some of the logistical ‘hiccoughs’ that were encountered in the simulation. There
were two unanticipated resources issues and an additional oversight by the researcher in the
training procedure.

The first resource issue was a misunderstanding about the timings of the room
bookings for the simulation. Initially, the plan was to train in one of the university classrooms
and then move to a computer lab after the morning break to complete the training and move
straight into the marking portion of the day. There was a miscommunication or
misunderstanding that led to the computer lab not being available at the time that the
researcher thought it would be, and so there was some disruption. It was recorded in the
researcher’s voice memos that this was regrettable but not a serious issue from his
perspective.

The second resource issue was with four participants not being able to log in to the
marking system. In the researcher’s notes, he recorded that the markers were not able to
access the marking system until approximately 1.15pm or 1.20pm, meaning that the
participants missed approximately one hour of online marking before the lunch break and a
SIMULATED LARGE-SCALE MARKING

small amount of additional time afterwards. The researcher had these participants mark the practice scripts in the back of the marker training booklet manually in the intervening time so that the participants were still working with student work and the marking guide, but unfortunately not in the same way as their fellow participants.

The most serious issue, which the researcher attributed in one of the voice memos to the disruption around the room bookings, was that a planned moderation session was overlooked, which meant that directly after the marker training the participants moved straight into marking. This was a mistake on the researcher’s behalf, as it did not follow the plan for the day. The moderation activity that had been planned to follow the marker training was instead held in the afternoon, at approximately 2.30pm, after a short break. In the researcher’s field notes, he observed that the discussion in this session was good, and that it covered strategies the participants had developed themselves for applying the relatively large number of criteria in a specific order to make the marking easier. He also noted that there was still some variability in judgments in the consensus marking and noted that this might be caused by insufficient training. However, in light of the reasonable interrater reliability reported earlier in this chapter, perhaps this perception of variability was not grounded in fact but rather in the researcher’s expectation or other cognitive bias.

Discussion

This chapter has presented the results of the analysis of three sources of information about the marking simulation: the likeness of the final scores assigned by the participants to the final scores assigned in authentic marking; an examination of the relative severity of the student markers in comparison with the ACER marking panel; and lastly, the field notes and reflections of the researcher, who facilitated the simulation. As reported in the results section, the quantitative analyses suggest that the simulation participants were able to reliably apply the marking guide to the student work, and generally were not very different in severity to the collective ACER marking panel. The qualitative analysis revealed that there were some aspects of the marking simulation where the fidelity of the simulation may have been compromised, and some differences between the simulated and authentic marking operations were noted.

The central question that this analysis is seeking to answer is the fourth guiding research question: “Can large-scale marking be effectively simulated?” As noted in the introduction to this chapter, there are several senses in which a simulation may be considered effective. For some, the key to an effective simulation is the faithful replication of a real
experience for the participants (Girod & Girod, 2008). This is true in some other fields of education, such as nursing education, where the fidelity of the simulation is commonly a primary concern; however, its importance and the optimal level of fidelity is contested in the literature. There is apparently no consensus as to whether the level of fidelity has a measurable impact on outcomes, although it is reported that generally participants’ attitudes are more positive when the simulation has a higher level of fidelity (Adamson, 2015). According to Hopwood (2017) the fidelity of a simulation is not a primary criterion of success or effectiveness; rather, because of the pedagogical intent of simulations, the learning outcomes are the primary concern and fidelity is a secondary or minor concern. The presence of a facilitator who is able to comment on and ask questions during the simulation is one way in which a simulation can differ considerably from a real situation and yet, according to Hopwood (2017), this difference supports the educational aims of the simulation. As Adamson noted, there is a lack of evidence that different levels of fidelity in nursing simulations affect their learning outcomes (2015). In many ways, Hopwood’s perspective, which de-emphasises the importance of fidelity, is in keeping with this (Adamson, 2015; Hopwood, 2017).

In attempting to establish whether this marking simulation effectively simulated large-scale marking, two further questions can be asked of it. Firstly, to what extent was the simulation an accurate representation of a real marking session? And secondly, where there were compromises in fidelity, were the differences likely to have supported the educational aims of the simulation, or were they more likely to have had a deleterious effect on the intended outcomes? The first question takes the conventional approach to simulation in expecting that there be a reasonable amount of faithfulness to a typical or ‘average’ marking experience. The emphasis in the second question is less on the fidelity of the reproduction, but the consequences of deviations from the conceptualised ‘typical’ large-scale marking experience. By exploring the answers to both questions, this research is considering the simulation from multiple perspectives, and in doing so, aims to understand more broadly what may have been accomplished. In short, the answers to these questions shed light on whether it is likely that large-scale marking can be effectively simulated at all, and whether the simulation in this research was able to do so.

As noted before, the simulation was largely based on a marking experience managed by the researcher in the course of his employment at ACER. This placed the researcher in a position to make observations about the similarity between the simulated marking experience and the specific instance of large-scale marking that this research attempted to simulate.
Rather than attempting to conceptualise a general idea of large-scale marking, or to attempt to create an experience that represented a ‘typical’ experience, the simulation in this research was modelled on a single instance. This allows for the researcher’s recollections of the real experience and the simulation to be compared in a way that is not always possible when simulations draw on a generalised version of a practice or scenario.

In this regard, the researcher’s field notes and memos are valuable, because the comparisons they contain provide evidence about the extent to which the simulation was able to recreate the real marking experience. In these sources, the researcher noted that the participants’ attitudes to marking were different to those of the professional markers who were employed in the original marking operation. Namely, the extent to which the participants discussed marking decisions and the qualities of the scripts with those near them differed between the two operations. This was a significant difference, but there is a question as to what extent this is likely to have affected the outcomes. This particular aspect of the simulation experience is addressed more fully in Chapter 6, where the participants’ perspectives are presented and discussed. Here, it will suffice to point out that the difference was noted by the researcher very early in the marking portion of the simulation and a deliberate decision was made not to interfere. Whether this decision was correct is a matter of speculation, but had there been more quantitative data – more scores assigned by the participants – it would have been possible to examine the extent to which the discussion may have affected the participants’ judgments.

In examining the reliability of the markers, the marks each assigned were compared with the scores that resulted from the original marking operation. So, the examination of interrater reliability was not done by comparing novice markers (the simulation participants) with other novice markers, but by comparing the novice markers with expert markers. The percentage agreement rates and the correlation in total scores indicates that the simulation participants were reliable in their judgments according to the commonly used heuristics for these metrics (Brown et al., 2004). A many-facet Rasch analysis of the participants’ severity also showed that there was not much difference in how harshly most participants and professional markers rated the responses. From these more quantitative perspectives, the amount of discussion in the marking room during the simulation appears to not have been detrimental to the quality of the marking. In other words, the ‘inauthentic’ discussion in the marking room during the simulation appears not to have affected the accuracy of the scoring. An argument can be made, then, that the marker training and processes were effective to the extent that there was little difference between the product of the simulation marking session
SIMULATED LARGE-SCALE MARKING

and that of the original marking session. So, when considered from that perspective, one answer to the guiding research question “Can large-scale marking be effectively simulated?” is that this simulation was able to effectively train preservice teachers to score reliably and accurately, and therefore it is likely that large-scale marking simulations can be effective in this respect.

The researcher’s perspective also offered comments on some logistical issues with the simulation: a misunderstanding with room bookings and an oversight with the scheduling of a moderation activity. On the one hand, these issues are unique to the simulation and diminish its fidelity because they were not encountered in the original marking operation. On the other, such issues are relatively minor. Additionally, it will be seen in the next chapter that the participants did not remark upon the room-booking issue in post-simulation interviews and emails, suggesting that perhaps it was unimportant to them. However, one participant commented that they wanted more moderation sessions. Perhaps, then, the scheduling issue did affect the experiential outcomes of the marking simulation, even if it did not appear to affect the scoring quality. That evaluation relies on the participants’ perspectives of the simulation, which are presented and discussed in Chapter 6, rather than on the researcher perspectives presented in this chapter. The participants’ perceptions notwithstanding, from the researcher’s perspective these compromises in fidelity were minor.

As said above, one of the ways to approach the guiding research question being focused on here is to consider the extent to which the simulation was able to faithfully recreate the authentic marking operation. The pragmatic answer, when considering the quantitative outcomes of the simulation and the researcher’s field notes and reflections, is that the training delivered in the simulation is likely to have been reasonably faithful to the authentic experience, because of the similarity of the end ‘product’: participants’ scores were found to broadly agree with scores from professional markers, and there was no evidence of an overall difference in severity between the participants and the professional marking panel.

That the ends of a process can be used to judge the process itself is, of course, a utilitarian argument. This approach to answering the guiding research question is acceptable within a pragmatic research paradigm because it is convenient, efficient and straightforward. However, this pragmatic, utilitarian argument largely ignores the experiential processes of the simulation. This point is revisited in Chapter 9.
Chapter Summary

This chapter addressed the fourth guiding research question of whether large-scale marking can be effectively simulated. Using the scoring results and the researcher’s field notes and observations, an affirmative answer was given: the results of the simulated and authentic marking were quantitatively similar, and the researcher’s impression was that any variations from the authentic marking experience were inconsequential to the simulation’s fidelity. However, this answer is based largely on the researcher’s perspective. In Chapter 6, then, the perspectives of the simulation participants are considered, to explore further whether the simulation adequately replicated the authentic marking operation upon which it was based.
Chapter 6 – Participant Experiences of the Marking simulation

While Chapter 5 offered the researcher’s perspective of the marking simulation, this chapter focuses on the participants’ experiences. In order to gather evidence from which to infer participants’ perceptions of the simulation, the researcher interviewed as many of the participants as possible beforehand, asking for email responses from those who were unable to meet. He also conducted small-group interviews directly after the simulation, encouraging email feedback from participants who did not take part in a group interview.

The perceptions of the participants are important for the line of enquiry summarised by the fifth guiding research question: “What do participants in a simulation of large-scale marking report about the experience?” In order to answer this question thoroughly, the chapter begins with a more detailed explanation of how the data were collected than was given in Chapter 3, and reiterates the data analysis methodology that has been used here and elsewhere in this research: the Interactive Method (Miles et al., 2014). Next, the results are reported in two parts. The first part reports on the analysis of the pre-simulation data, identifying themes related to participants’ experiences of marking and reasons for wanting to participate. The second part presents the analysis of the post-simulation data, identifying themes related to the participants’ experiences of the simulation, including their perceptions of the day’s value and of any knowledge, insights or skills they gained or developed while taking part. The chapter closes with a discussion that synthesises the two parts of the data analysis, to concisely but comprehensively answer the question of how the participants perceived the simulation shortly before and immediately afterwards. Further, implications for any future iteration of this marking simulation that extend from the participants’ perceptions of this simulation are considered.

Methods

Data Collection

Prior to the day of the simulation, interviews were able to be arranged with 13 of the participants. The aim of the interviews was to gather data about what, if any, expectations or preconceptions the participants had about the simulation, and what previous experience they had with marking student work. These interviews were semi-structured, using a set of guiding questions (Appendix H) to ensure each interview covered similar ground, while remaining open to investigating observations, issues and comments that were raised in order to achieve
a greater level of investigative depth. The typical length of the interviews was five to seven minutes, though one lasted three minutes, and one ran slightly more than 12 minutes.

All interviews were recorded, with two recording devices being used: a laptop and a Livescribe Echo smartpen. A smartpen is a recording device that records audio while also digitising the user’s notes as they are written. This results in a data file that displays the notes on a computer as a ‘clickable’ document, where selecting a particular word or mark on the page cues the audio recording to the time that text was written. All interviews except one were recorded on both devices simultaneously as a failsafe in case of user error, one device failing completely or a failure to record adequately good quality audio. There was a single interview where only the smartpen was used, as the interview was conducted in a location not conducive to setting the laptop up to record.

The audio recordings were reviewed after all interviews had been conducted, and one of the two recorded versions was selected as having higher quality sound. New copies of those files were made, with the originals being archived, so that excess silence and background noise at the beginning and ending of the recordings could be trimmed from the files before they were uploaded to an online transcription service. The transcription service Way With Words (waywithwords.net) was selected in consultation with the researcher’s primary supervisor. This decision was based on feedback from that supervisor’s other students, who had found the service reliable and cost-effective. Eleven of the 13 interviews were sent to the transcription service.

Two interviews were not sent to the transcription service. These two interviews were longer than most, both being more than 10 minutes, and both recordings of each had considerable background noise, making portions difficult to hear. These interviews were longer than most because the interviewees had wide ranging thoughts about assessment and teaching that went beyond the guiding questions, but which were still relevant to the general aims of the interviews. Due to the deviation from the typical structure of the other interviews, and the poorer quality of the audio, the researcher decided to transcribe the relevant parts of these interviews himself.

In all cases, the focus of the transcripts was the content and ideas contained within, rather than the speech or interactions between interviewer and interviewee, and so denaturalised transcription was suitable (Oliver, Serovich, & Mason, 2006). In this approach, unintended vocalisations, extended pauses, self-corrections, and so on are deprioritised in favour of transcribing the content of the dialogue. The transcription still aims for accuracy,
SIMULATED LARGE-SCALE MARKING

but as Olivier et al. (2006) wrote, “Accuracy concerns the substance of the interview, that is, the meanings and perceptions created and shared during a conversation.”

A number of volunteers were unable to meet in person for an interview, so as an alternative they were invited to email the researcher with information about what marking they had done previously and what outcomes they hoped for. The email was worded as follows:

If you are unavailable to meet me before Wednesday, then I’d still like to hear back from you by email. I’d very much like to get a few thoughts from all participants about the following:
What marking you’ve done before (on prac, perhaps?).
How you feel about marking.
Why you volunteered to participate on Wednesday.
What you hope to get out of Wednesday’s session.

Email responses were collected from all of the remaining nine participants, meaning that data from all participants was collected prior to the simulation.

Post-simulation, two small-group interviews were conducted and again were recorded using a laptop and a smartpen, using a set of guiding questions that are also included at Appendix H. In the first small-group interview there were six participants, and seven participated in the second interview. In total, more than half of all participants were able to contribute their perceptions of the simulation face-to-face. In emails after the simulation, the researcher asked participants if they had any additional feedback, and if so, to reply by email. Additional email feedback was received from three participants, two of whom had participated in the small-group interviews, and one who had not.

Data Analysis

As with the analysis in phase one, the methodology for the analysis of the qualitative data followed the Interactive Model of data analysis as outlined by Miles et al. (2014). Two NVivo projects were created: one for the pre-simulation data, and one for the post-simulation data, and so the data were analysed independently of one another. All interview transcriptions and audio, along with relevant email extracts, were loaded into the relevant project. The pre-simulation project was created and coded first, then the post-simulation project. This division was made in order to separately identify the themes present in the data sets before comparing them.

To maintain participant anonymity, each participant was assigned a unique identifier. In the rest of this chapter the participants are referred to only as Marker n, where n is a number from 1 to 22. The numbers were not assigned alphabetically, nor in any other
systematic way, in order to dissociate the identifier from any attribute of the participant and to preserve participants’ anonymity. The researcher has also tried to not refer to markers by pronouns that imply sex or gender.

Coding of the data was performed by only the researcher, who followed the processes of the interactive model of qualitative analysis (Miles et al., 2014) in the same way as in other parts of this research. The methodology is explained in detail in Chapter 3.

In some cases, a comment from a participant, whether in an interview or in an email, was coded to multiple nodes because it was interpreted to be relevant to more than one theme or sub-theme. For example, one participant commented in an interview,

I’ve done none which is why this really interests me because … one of the things that I’m really worried about is marking and being … not consistent … I think I could be consistent. I’m more concerned about what level we’re marking at, you know what I mean? Are my expectations going to be too high, or are they going to be too low? (Marker 6)

This comment is multilayered: the participant provides information about having no past marking experience and expresses a lack of confidence in their own marking ability. They speak about hoping to achieve consistency in judgment, and also consistency with broader, perhaps shared, expectations of students’ work. This is just one example of the richness of the data – in only a few sentences this participant has been able to express not only their goals for participation, but also provided information about their past marking experience and how they feel about marking. Each of these ideas or themes deserves separate recognition, and so it was appropriate for parts of the extract above to be coded to multiple nodes in order to attempt to capture the complexity of the response.

Results

The results of this phase of data collection are presented below in two parts. Firstly, the results of the analysis of the pre-simulation interviews and email responses are reported. Then, the results of the analysis of the post-simulation, small-group interviews and email feedback are reported. The chapter ends with a comparison of what the preservice teachers reported about the simulated marking experience and what teachers reported about participating in authentic marking, pointing out the similarities and differences between each group’s perceptions.
Pre-Simulation Interviews and Email Responses

There were two first-level nodes created in the analysis of the pre-simulation data, each one representing a major theme present in the data that provides some sense of the participants’ pre-simulation attitudes towards and perceptions of marking and the marking simulation. These first-level nodes were further broken down into sub-themes. Table 12 below displays the themes and sub-themes identified in the emails and interviews conducted prior to the marking simulation.

Table 12 Themes in pre-simulation interviews and emails.

<table>
<thead>
<tr>
<th>Experiences of marking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lack of certainty</td>
</tr>
<tr>
<td>Quantity or amount</td>
</tr>
<tr>
<td>Stress and time</td>
</tr>
<tr>
<td>Harshness</td>
</tr>
<tr>
<td>Goals for participating</td>
</tr>
<tr>
<td>Gain experience</td>
</tr>
<tr>
<td>Gain confidence</td>
</tr>
<tr>
<td>Efficiency</td>
</tr>
<tr>
<td>Setting expectations</td>
</tr>
<tr>
<td>Professional development</td>
</tr>
<tr>
<td>Increased objectivity</td>
</tr>
</tbody>
</table>

In reporting the results of this phase of data gathering, the responses coded to each theme and its sub-themes are compared within each theme and the responses are generally characterised. These characterisations are then synthesised in a summary section to generate general statements about the simulation participants’ goals, expectations, and perceptions of marking and the marking simulation.

Experiences of marking. Each interview, and the first question in the email, was intended to elicit from participants what experience they had of assessing student work. This information was collected as much to establish what prior knowledge could be assumed in the training workshop as it was to determine whether the amount or quality of participants’ past experiences with assessment contributed to their decision to participate. As will be seen later in this chapter, the key themes identified in the responses about past marking experience appear to be echoed in participants’ responses about their goals for marking.
Lack of certainty. Within the references coded to ‘Experiences of marking,’ 10 participants spoke about being uncertain in making judgments about student work. A very direct example was from Marker 3’s email: “I don’t feel confident at all that I can evaluate a piece of student writing properly. And by properly, I mean in a way that would meet the consensus of the piece’s quality reached by my peers.” Variations of statements like this were present in other responses: there were 12 similar references coded in the responses from the 10 participants. Marker 17 directly linked uncertainty with lack of experience with marking, saying, “I don’t feel very confident with marking in general as I don’t feel I have had adequate experience.” This link between uncertainty and the sub-theme of ‘Quantity or amount’ is intuitive. It appears reasonable that a lack of practice and familiarity with judging students’ work would lead to uncertainty in one’s own judgments. It would naturally follow, then, that more experience with marking would help to build confidence in judging student work. This aligns closely with the overarching aims of this research, so is revisited later.

Quantity or amount. The participants recruited for the marking simulation were all education students in the final year of their course. However, 14 of the 22 participants were enrolled in a Graduate Diploma of Education (DipEd) course that had a one-year duration, so the last year of their course was also the first year. Unsurprisingly, given the simulation was approximately half-way through the course, these participants had limited experience teaching in schools and consequently limited experience with assessing student work. That said, some had experience with assessment in other contexts through private tutoring, overseas English as a Second Language teaching, or tutoring at a university, for example. Additionally, all participants except one had undertaken at least one practicum, so most had some marking experience, though the extent of this experience varied considerably. As an illustration of this variability, Marker 1 had a very limited experience with assessment on practicum: “I didn’t do a lot of marking for it, but I remember, my mentor teacher did give me like one Year 12 exam to like have a practise at, and it was really hard to like make decisions.” Another participant commented (by email), “I have only done marking in the past of multiple choice – which has set answers” (Marker 11). In contrast to these responses, Marker 2 responded in an email that

I was involved in moderating, cut-off discussions and curriculum marking expectation meetings. I also marked 40 Yr 7 creative writing projects, 40 Yr 9 English extension short answer assessments and 40 Yr 9 mainstream English short answer assessments as well as 40 essay plans. These were done after a short ‘moderating’ meeting with my mentor where we went through 2 or 3 together before I went off and marked alone.
SIMULATED LARGE-SCALE MARKING

All three participants were enrolled in DipEd courses, and had completed one of the two professional experience placements required for their course. Thus, the marking experiences of participants were idiosyncratic to say the least – two had almost no practical experience of assessment, while the other had already been included in professional practices such as moderation and standard-setting meetings, as well as cross-marking.

Even among the participants who were undertaking a four-year course, where students are required to complete professional experience placements each year of the course, with a 10-week (one Australian school term) placement in the first semester of their final year called the Assistant Teacher Program (ATP), experiences varied and seemed to rely largely on the particulars of the practicum placement. Marker 4 gave a reasonably long response by email that encapsulates several ways that practicum experiences can vary, not only between students, but between practicums for an individual student:

My experience with marking was almost non-existent until my ATP. On my 5 week prac, I marked alongside my mentor teacher for Year 10 Oral presentations; however, on my ATP I was solely responsible for marking everything.

I marked 2 classes of Year 9 informal viewing assessments, 2 classes of Year 9 informal speaking and listening assessments, 1 class of Year 8 informal viewing and 1 class of Year 8 informal speaking and listening.

I also marked 2 classes of Year 9 final Essays, 1 class of Year 8 final essays, 1 class of Year 7 final essays, 1 class of Year 8 final creative assessment, and 1 class of Year 7 final Speaking and Listening assessments.

I found the experience stressful and not as rewarding as I would have hoped. There was a very basic rubric, and I was repeatedly told that marking was ‘teacher judgement’, which I don’t actually believe, and nor do I think I have the experience as a teacher to be qualified to even make a ‘teacher judgement’. None of my marking was moderated, so I went over each piece multiple times to try and be as correct as possible, which means it was both time consuming and pointless – as I still don’t know whether my marks were even correct!

I found the allocation of marks to be completely arbitrary: “decide whether it’s an A, B, or C, and then choose a mark that reflects that,” so my experience was not very positive. (Marker 4)

This response touches on many themes. It speaks not only of the quantity of marking done by the student teacher, but also characterises the marking experience as stressful, expresses their lack of confidence in making consistent or ‘correct’ judgments, and raises doubts about whether ‘teacher judgment’ is a valid basis for assessment. Reports of and references to negative emotions connected with marking appeared in several responses. Some appeared often enough to warrant the creation of new nodes.
**Stress and time.** During coding, ‘Stress’ and ‘Time consuming’ were initially identified as separate sub-themes. However, in further iterations of data condensation, following the principles of the Interactive Model (Miles et al., 2014), they were conflated. The two themes were merged because both are negative experiences of marking that are likely to be interrelated: time pressure is stressful, and the students referring to stress are doing so within the context of having to meet feedback deadlines or feeling pressure about making correct decisions. However, the theme is minor in the dataset: only three participants referred to stress, and a different three referred to the time-consuming nature of marking.

One participant who had done more marking during practicum than most spoke of “slogging through English assessments … which, due to the sheer volume of work, writing, etc. often took me well over 4 hours to fully mark one class” (Marker 2). The hours spent on this marking would presumably have been either between other teaching duties or, more likely, out of working hours in the evenings or early mornings. That Marker 2 refers to marking as ‘slogging through’ implies that there is both drudgery and difficulty in the task. The time-consuming nature of it may, with a small amount of extrapolation, contribute to a preservice teacher’s feeling of stress or pressure – a further reason to consider these two sub-themes together rather than separate.

Marker 18 spoke about being intimidated by the time-consuming nature of marking, as well as the pressure of students wanting a quick turn-around:

Marker 18: My teacher, my mentor teacher, actually had year 11 and year 12 exam papers that they had done that the week before I got there, and one class had 27 exams with three essays in each one. And they were already asking when they were getting their marks in and she’s just like, “it takes me an hour to mark one person’s three exams.” So yes. It scared me.

Researcher: Scared you? You’re daunted by marking one?

Marker 18: Because I don’t have that much time yes.

In this response, the stress and fear result from the participant’s observations and conversations with another teacher, not personal experience. However, that the participant’s anxiety did not come from personal experiences is no reason to disregard it.

That the professional responsibility of marking takes a considerable amount of time, and that it adds to a teacher’s stress, is far from surprising. Nonetheless, making judgments about students’ work is part of a teacher’s professional duties, as made clear in the Australian Professional Standards for Teachers (Australian Institute for Teaching and School Leadership (AITSL), 2011). Encouragingly, one participant said of marking that “I found the experience stressful but rewarding” (Marker 13), and another wrote that on practicum,
To begin with I found marking quite stressful. I doubted my decisions and was wary of being too harsh or too lenient. However, after mediating with my mentor teachers and finding out that my marks where on par with their own marks I felt more confident. (Marker 8)

In other words, with more practice, and conversations with colleagues and mentors, Marker 8 grew in confidence, thus reducing the stress of marking. For that participant, at least, practice and experience appeared to be effective in ameliorating the negative emotions associated with marking.

**Harshness.** Three participants mentioned that they found themselves to be more harsh or severe in their marking than their mentor teachers. Take Marker 14’s reflection that what I’ve struggled with is marking according to their grade level because it was year eight and I wasn’t sure what to expect from year eight. And the feedback that I got from my mentor was you’re marking a bit harsh … So I struggled with that.

Again, this participant appears to associate negative emotions with the marking experience: they characterise applying fair or reasonable expectations in assessing students’ work as a struggle. This has strong parallels to the negative emotions discussed in the previous sub-theme of ‘Stress and time,’ but it is different because it relates more to the difficulty of making accurate and reliable judgments, not just to the pressure of expectation.

**Summary of key themes on the topic of marking experience**

This section has reported on the high amount of variability in participants’ previous experiences of marking on professional placements. This is in keeping with literature that highlights the idiosyncratic nature of teacher practicum experiences (Rorrison, 2008). Each participant appeared to have had different experiences on practicum: some had large marking loads and a lot of responsibility given them by their mentor teacher, while others had almost none. In addition, this section identified several minor themes relating to preservice teachers’ negative emotions about marking. The most common negative emotions reported were feeling uncertain about judgments made about student work and concern over marking too harshly, and there were also a few references to stress and the time-consuming nature of marking.

**Goals for participating**

All 22 participants gave responses about the goals they had for their participation in the simulation, either in interviews or by email. Often, these were direct responses to a question from the interviewer about why they had volunteered, but there were occasions where participants expressed what they hoped to get out of the simulation in response to a
different question, or while responding to probing questions. References that were coded to ‘Goals for participating’ were further classified into six sub-themes. The sub-theme with the most identified references was ‘Gain experience,’ with references identified in 11 participants’ responses. Two more sub-themes were identified in at least five participants’ responses: ‘Gain confidence’ and ‘Efficiency.’ ‘Setting expectations’ was a sub-theme that was coded in four participants’ responses, and an additional two sub-themes were coded in at least two responses. These minor sub-themes are considered together below.

**Gain experience.** There were 11 participants who referred to gaining experience as being a goal for their participation in the marking simulation. Nine of the 11 participants whose responses were coded to this node were students in the one-year DipEd. course. In speaking about their goal of gaining experience, the participants demonstrated that they were quite aware of their lack of experience in marking and making judgments about students’ work. One participant saw this lack of experience as being related to the content of the course: “Just to like get more experience doing because I feel like in this course, the Grad.Dip, you’re usually just doing a lot of planning and that’s what you’re marked on. No one really marks your marking” (Marker 1). Another student in the DipEd course emphasised that the lack of experience was more to do with the short duration of the course, rather than the content of it: “I don’t have a lot of experience with marking and with the DipEd programs, you know, it’s a year-long program, so I’ll be a teacher next year, so, any experience would be great” (Marker 21).

One of the DipEd students was undertaking the course part-time, so had not yet had a practicum experience (part-time students only do practicums in their second year). They were therefore less concerned about entering teaching with a lack of experience as about going into a school as a teaching student with no experience:

Before I go and do prac I kind of wanted to have a little bit more knowledge around marking in different types of work that I might see or come across, so I don’t feel like a total ‘zilcho’ when I walk into a classroom. (Marker 5)

Given that these preservice teachers were enrolled in a relatively short course, it may not be surprising that they had concerns about their own level of experience and preparation and felt nervous as to whether they would be ‘classroom-ready.’

Statements such as, “I just think any experience I can get at doing it, is good, and if I can do it in a block then it will sink in” (Marker 6) can be interpreted as showing these students’ awareness of their inexperience. This inexperience, though, was perceived as a challenge to be solved, and participants indicated that they were keen to take up opportunities
SIMULATED LARGE-SCALE MARKING

to compensate for their self-identified lack of experience. Marker 12, for example, said “Because I don’t have a lot of teaching experience [and] I have no, like, background in education, [I’ll do] whatever I can do to upskill” (Marker 12).

The three students in the four-year Bachelor of Education (BEd) courses who referred to gaining experience as being a goal for participation expressed similar ideas to the DipEd students, despite having had more practicum experiences with marking. The most notable difference was that the BEd students recognised that they already had some experience, and that they were hoping to build upon that base by participating in the simulation. Marker 13 put it most succinctly in an email response: “I wanted to participate to gain some more experience with marking and to develop my skills (making fair and accurate judgements).” Even the participant who had reported doing the most marking of all participants still responded that “I am hoping to gain some more confidence and experience in my marking, which will make the first year of teaching slightly easier” (Marker 4). Perhaps the best summary of the sentiment of BEd students who reported gaining experience as their goal for participating comes from Marker 16: “I feel like marking’s one of those things where the more you can do, the better.”

In summary, many of the participants responded that a goal for participating in the marking simulation was to gain more experience in marking. Given that many of the participants had little previous experience of applied marking (but all having encountered the theory of assessment in their courses), this struck the researcher as an unsurprising motivation. How widespread this lack of practical experience is for preservice teachers in Australia is unknown: Grainger and Adie (2014) reported that they were unable to locate any Australian literature “that investigated how preservice teachers learn about and are inducted into key assessment practices” (p. 90), and the literature review in the present research, reported in Chapter 2, similarly identified this gap.

**Gain confidence.** Eight of the participants, some of whom had indicated they had some marking experience and others who had reported having little experience in marking, said that a goal they had for participating was to gain confidence in marking. In fact, the comment from Marker 4 that was quoted more fully just above mentions both experience and marking in the same sentence, and expresses that this marker was “hoping to gain some more confidence and experience in my marking.” Later in the interview, Marker 4 went on to say, “I just don’t feel confident about making decisions, even though I consistently was making similar decisions to my mentor teacher and the other teachers that I cross-marked [with],” and that “as soon as I was left to my own devices, panic set in.” Marker 20 also mentioned
experience and confidence together, saying, “I think just more confidence, more experience, and some validation, maybe, hopefully.” This is the only reference by a participant to ‘validation’ in this context, but it was interpreted to relate to the idea of confidence, and that Marker 20 was expressing a desire to receive external confirmation that their marking ability was adequate in some way.

Not all participants made an explicit connection between gaining confidence and additional experience, but in the broader context of the conversations and email responses, the researcher inferred that connection. It would appear to be reasonable for preservice teachers to assume that more practice would lead to gaining both experience and confidence – so much so that the researcher considered collapsing this node and the previous one into one theme. This was not done, primarily because the responses suggested that although they are connected, at least some of the participants saw the two goals as separate enough to mention both, even if in very close proximity.

Efficiency. One of the themes that the researcher had not expected in the responses about goals for participating was the participants’ desire to increase their marking efficiency. Seven participants referred either directly or indirectly to increasing their marking efficiency. Marker 1, for example, indirectly referred to efficiency, saying that a goal was to “make … quicker decisions and quicker judgements.” Marker 12 saw efficiency as a goal to be achieved, a view informed by an experience on practicum:

And why did I volunteer to get involved in it? Because there’s obviously a lot more to it than, you know, if you want to get the stage where you’re marking a couple of hundred papers in a week, which is what was the case on prac. One of the lads had done 350 in a week, ATAR … Or 250 ATAR Year 12, you know, examinations in a week, you know, that’s where you want to be going, as opposed to like, you know, five essays in a week. (Marker 12)

This response was not interpreted to be only about raw marking speed. Marker 12’s comment about there being “obviously a lot more to it,” along with some other context, indicated that they saw marking speed as being about skill and efficacy rather than about just getting the work done. Marker 10 hoped that the marking simulation would provide “a better, more effective system of marking,” which the researcher interpreted as an indirect reference to efficiency, and Marker 8 was “hoping to get practice in order to be more accurate without spending hours on the same piece of work.”

Whether the reference was direct or indirect, these seven participants saw the marking simulation as an opportunity to develop skills to achieve a goal of, in the words of Marker 14, “using my time more efficiently.”
**Setting expectations.** The researcher was surprised to find that only four participants referred to their expectations of students’ work. He had expected it to be a more obvious goal for undertaking the simulation, and that more participants would have had comments like

> I’m more concerned about what level we’re marking at, you know what I mean? Are my expectations going to be too high, or are they going to be too low? Where is that line, where is that baseline, and when they’ve got, maybe, different abilities, where does that baseline sit because [what if they’re] really great at grammar or the whole subject thing … but not particularly good at the practical side of things. How do you weight it? (Marker 6)

This participant explicitly voiced their concern about having reasonable expectations against which to judge students’ work. Two participants expressed an understanding that expectations are not necessarily an externally set standard, briefly commenting that their goal was about “improvement in my ability to mark to a consistent level to my peers” (Marker 9), or expressing a hope that “there’s some sort of comparison that I can see with what I’m thinking and then what other people, my peers are thinking” (Marker 18). Marker 14 hoped to develop skills in “how to recognise what to look for in different grade levels”. They clarified that “as an English major I am not sure how much I should be marking spelling, grammar etc or I guess addressing the task specifically.”

Standards and expectations are such a central part of assessment that the researcher was very surprised that there was not more mention of them among participant responses. Nonetheless, this theme was minor in the participants’ pre-simulation responses.

**Minor themes.** There were only two more themes that were identified in more than one participant’s response: PD and increased objectivity. These minor themes do not bear much discussion because they were only mentioned briefly and by few. However, PD was a theme that emerged in the research results of phase one, so it may warrant slightly more attention.

Two participants referred directly or indirectly to PD. Marker 3’s reason for volunteering was “I realise that marking is an essential skill for a teacher to possess. And right now it’s an area I need to improve.” Marker 7 observed that “professional development is a lifestyle for teachers.” Of course, the recruitment materials had stated that participants would receive a certificate confirming they had completed 6.5 hours of PD, so PD was an explicitly stated purpose of the marking simulation. In which case, of course, the absence of discussion about PD as a purpose or goal is, perhaps, even more curious. It may be that participants considered gaining experience and confidence as more important than some vague notion of PD.
Summary of key goals for participating

The goals participants had for volunteering for the marking simulation revealed three main themes. The first two were to gain additional marking experience and to gain additional confidence in making judgments about student work. A third, less frequently expressed theme concerned setting expectations about student performances. This third theme is related to having confidence in one’s judgments inasmuch as being more certain of standards or expectations can lead a marker to feel surer that he or she has made a correct judgment. Therefore, the key goals for participating can succinctly be expressed: students volunteering to participate did so to gain experience and confidence in marking.

Summary of Themes in Pre-Simulation Interviews and Emails

The results of the pre-simulation data analysis were broadly divided into two sections: what previous marking experiences the participants had, and what they hoped to achieve through participating. As it eventuated, the results of the two sections are closely related.

The amount of marking experience participants had varied wildly, ranging from no practical marking experience to responsibility for marking a quantity of student work equivalent to that which a full-time teacher might take on, for an entire term-long professional placement. Additionally, a few participants even had marking experience in contexts other than school education. A participant’s amount of previous marking experience, though, was not necessarily a predictor of whether they identified additional experience as a goal of participation – some of the participants with comparatively more experience also articulated a desire for more experience. So, in inquiring about participants’ previous marking experience, it was discovered that despite how much experience they had, they articulated a desire to gain more, and this was also one of the key goals for participating identified.

In addition to the goal of gaining experience was another goal that, really, goes hand in hand with it: increasing confidence. Gaining, building or increasing in confidence was the other main theme in participants’ responses, and many of the participants reported that one of their main reasons for volunteering was to become more confident in making judgments about student work. In the perceptions that participants reported in the interviews and emails, it was clear that participants saw these two goals as related as well.

The conventional wisdom is that the more one does an activity, the more one increases in competence, and the more one feels confident. This observation would go quite some way to explaining not only what reasons the participants had for volunteering for the simulation, but also what the participants aimed to achieve by participating. Bluntly, the
participants felt that they had not marked enough and so they lacked confidence in their marking ability; by volunteering to take part in the simulation, they hoped to gain more of both.

Post-Simulation Interviews and Emails

Post-simulation data was collected from only 14 of the 22 simulation participants. There were two group interviews: one in the last hour of marking, and the other immediately after the simulation. There were six participants in the first group interview, and seven in the second. Additionally, all participants were invited to email any additional thoughts they had about the day. One additional participant provided a response by email, while two others who had participated in a group interview provided additional thoughts in emails in the days following the simulation.

There were four first-level nodes created in the analysis of the post-simulation data, each node representing a major theme present in the data that contributes to understanding the participants’ impressions of the simulation. Additional themes were identified: sometimes these were identified within the material already coded to a theme, and other times they were at first identified as a separate theme and later incorporated into one of the major themes. The organisation and reorganisation of themes and sub-themes is allowed for in the Interactive Model of qualitative analysis (Miles et al., 2014), which conceptualises qualitative analysis as moving between phases of data collection, data display, data condensation and drawing conclusions. Reorganising the hierarchy of themes during analysis is data condensation. Table 13 shows the final arrangement of the themes identified in the analysis.
Table 13 Themes in post-simulation group interviews and emails.

<table>
<thead>
<tr>
<th>Outcomes and achieved goals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rubrics</td>
</tr>
<tr>
<td>Confidence</td>
</tr>
<tr>
<td>Experience</td>
</tr>
<tr>
<td>Ideas for future teaching</td>
</tr>
<tr>
<td>Exemplification of standards</td>
</tr>
<tr>
<td>Objectivity</td>
</tr>
<tr>
<td>Perceptions of large-scale marking</td>
</tr>
<tr>
<td>Difficulties</td>
</tr>
<tr>
<td>Noise</td>
</tr>
<tr>
<td>Duration</td>
</tr>
<tr>
<td>Uncertainty</td>
</tr>
<tr>
<td>Future participation</td>
</tr>
<tr>
<td>Marking experience as part of ITE</td>
</tr>
<tr>
<td>Narrative marking</td>
</tr>
<tr>
<td>Feedback for future iterations</td>
</tr>
</tbody>
</table>

The group interviews were semi-structured, with a set of five guiding questions. So, some similarity in the themes between the two groups would be expected. Figure 10 graphically compares the small group interviews by representing which cases and nodes are connected to each of the two interview transcripts. The participants in each small-group interview are identified by the marker number that has been used for them in all parts of this research, so cross-referencing is possible. The majority of themes are in the centre of the figure because they are common to both interviews. Both of the small-group interviews have sub-themes unique to each, with ‘Objectivity’ and ‘Perceptions of large-scale marking’ being unique to Small-Group Interview 1, and ‘Narrative marking’ being unique to Small-Group Interview 2.

There were also some emails from participants received after the simulation day stating further impressions. Emails were received from Markers 9 and 12, who had each also participated in a small-group interview, as well as Marker 6, who had not participated in an interview. There were four themes in the emails: Marking experience as part of ITE; Noise; Future participation; and Difficulties. All four of these themes were common to the two interviews as well, and no themes were unique to the emails.

In reporting the results of this phase of data collection, the four major themes are presented along with synthesis and generalisation of the sub-themes and references associated with them. Where a sub-theme is unique to only one of the two group interviews this is noted.
but it is not considered in isolation from the other data in the theme. The decision to present the results like this was made because there is more in common between the two interviews and the emails than there is different. Despite the fact that the data collection happened at different times and mostly involved different participants, all data sources present views about the same central event and so the researcher decided to treat the data as a single set of perceptions and experiences.

In the final part of this section, the experiences and perceptions gathered from the participants are synthesised to generate general statements about the simulation participants’ impressions of the marking simulation. Then, in the final section of the chapter, the pre- and post-simulation perspectives of the participants are further compared and generalised.

**Outcomes and achieved goals.** This theme is a grouping of responses that expressed something about the participants’ ‘takeaways’ from the simulation – the lessons they feel they learned from the simulated marking experience. In some cases, the participants spoke about achieving the aims and goals they had prior to the simulation, while in others the participants expressed something new or surprising that they encountered during the
SIMULATED LARGE-SCALE MARKING

simulation. There were seven sub-themes identified within this theme; they are each discussed below in order of the number of references coded to each.

**Rubrics.** This was the sub-theme with the most references across both small-group interviews related to what the participants had learned from the simulation, though rubrics were not mentioned in the post-simulation emails from participants. In the interviews, several participants commented on and discussed the marking rubric used in the simulation. As stated earlier, the rubric was developed by ACER for its *eWrite* assessment, the assessment that was the basis for the simulation. The participants, on the whole, reported that it was useful and easy to apply in making judgments, though one participant commented on finding a specific part challenging to interpret:

> With the rubric, there was one section … As I’m not an English specialisation teacher, there was a section. One of the, I think it was one of the punctuation ones. I struggled a little bit with the explanation in the rubric. There wasn’t enough for me to go off of. (Marker 11)

Despite that comment, the consensus appeared to be that the participants were satisfied in the quality of the rubric, which led several to comment on the usefulness of rubrics generally, and also to compare the rubric used in the simulation to rubrics they had encountered in their own education or on their teaching practicums. Marker 3 was one of the participants who spoke about the benefits of a good rubric in the first small-group interview, saying

> I think, having a well-written rubric, this has just kind of really impressed on me the importance of a good rubric. Obviously, to help your students write well, for them to know how they can write well. But, also, to focus your marking on more objective criteria and not just looking at an overall impression of the piece. And I think, also, if it did come to writing comments, the rubric really does give you meat for, then thinking, okay well how can this student improve? You can really look at those things under those different categories, where their strengths are and where their weaknesses are. (Marker 3).

It is notable that Marker 3 was not presenting ideas about the use of rubrics in large-scale marking but was generalising to future classroom practice. This was an encouraging indication that perhaps the simulation of large-scale marking could help preservice teachers in their professional practice after graduation – a key hypothesis of this research.

Markers 8 and 4, who participated in the second small group interview, offered a perspective on the advantages of having analytic marking criteria in a rubric during a short interaction:

> I had some marking experience before with, I would say, quite a loose rubric, and it just takes you double the time. When I first saw this rubric, I was stressed because it was like, it’s going to take more. But, in fact, it takes much less. When you have
categories and specific things you’re looking for in separate sort of ideas, it’s so much easier than stressing about one word in a lot of paragraphs. That’s quite loosey-goosey. So, the importance of a good rubric … (Marker 8)

And, I think we’re kind of taught that through our teacher education. I know we’ve been taught about how important it is to have this really differentiated rubric. But, out in schools, you just don’t see it. It definitely made it a lot easier, marking real pieces of work with a decent rubric, than being told, “oh it’s all about teacher judgment, just pick a number,” basically, and that’s where they sit, which is my marking experience in my ATP. (Marker 4)

Marker 20, who participated in the other small-group interview, expressed a similar idea, saying that “the rubrics I saw on Prac were nowhere near as easy to follow.” This was further articulated by Marker 20, who explained that having a number of specific, discrete criteria instead of more general criteria made judgment easier. Marker 20 said that in one rubric, “the paragraphing, grammatical use [and] sentence structure was all together” and so determining a score for a student who displayed strengths in one of these areas but not the other was quite difficult. Marker 20 suggested, “I’m wondering if the answer is, split them into more boxes?”

Again, Marker 20 was generalising the simulated large-scale marking experience to classroom assessment.

Another perspective offered was on the usefulness of the morning training, where the rubric was explained, and exemplars were matched to the criteria descriptors to illustrate the score points within each criterion. Marker 2 said of this experience,

One of the best bits about today is just taking the time to learn how to read and use a rubric. Because the biggest thing I heard when I was out [on practicum], was teachers begging for PD on how to read the rubrics they’d been given. How to read things coming from [the School Curriculum and Standards Authority] the way that everyone else is reading them. Is there something standard? And just being able to have a guide and not just being given something to interpret. I think just going, literally, just an hour and a half, to go, we’re reading through the rubric, this is what it means, and we’re going to show you what it looks like, was awesome. Just to be exposed to that. (Marker 2)

Marker 2 appears to be expressing a perception that teachers in schools who are given rubrics, rather than making them themselves, experience frustration in applying them because they are unsure of either how to interpret them or how to reliably apply them. Other participants in that group interview generally agreed with Marker 2 that receiving training on the rubric prior to marking was a valuable experience.

So, on the whole, the comments in the post-simulation data on the topic of rubrics can be generalised as being generally positive about the quality and ease-of-application of the specific rubric used in the simulation. Participants often contrasted this with how they had seen rubrics used by teachers on practicum, where they had observed or experienced a lack of
SIMULATED LARGE-SCALE MARKING

clarity and uncertainty in applying ‘pre-made’ rubrics to student work. Participants also contrasted the simulation’s rubric against poorly constructed rubrics they had encountered. In several instances, participants framed these comparisons in terms of how they might incorporate this experience into their future classroom work. This generalisation beyond the simulation context was the first evidence that the simulation may have contributed something to these preservice teachers’ development into professional teachers.

Confidence and experience. As was reported in the analysis of the pre-simulation data, the two main reasons participants had for volunteering for the simulation were to become more confident in making judgments about student work and to gain experience in judging student work. In the post-simulation data, these sub-themes were coded separately, but they are reported together because references to the two were often in close proximity, implying that the participants perceived a connection between them. For example, Marker 7, participating in the first of the post-simulation small-group interviews, reported, “I feel more confident, and I just wanted to be exposed to more marking, so this has definitely ticked those two boxes.” Marker 5 expressed similar ideas in a different way:

I wanted to feel confident before I went into Prac or something along those lines because I haven’t done a Prac yet. So, I was just trying to get involved in some marking and some practical work in any way, shape or form. (Marker 5)

For these participants, and a few others, the goals of gaining confidence and gaining experience were clearly connected.

However, not all participants saw confidence as linked to experience. For example, Marker 16 spoke of a different pair of goals: “I think my big one was more confidence and more guaranteed consistency. I think is what I was after and I definitely feel that I’ve achieved that during the day,” while Marker 1 was focused on confidence: “I definitely feel more confident in my judgements that I make.”

On the whole, the comments referred to above imply that those participants who had volunteered with the goal of growing in confidence and gaining experience felt, immediately after the simulation, that they had achieved those goals. This was also the general feeling among the other participants. This is an important finding for this research, as it helps show that the participants perceived the simulation to have had a positive effect on them.

Ideas for future teaching. In references coded to the sub-themes above, particularly in the Rubrics sub-theme, it is noticeable that some participants expressed ideas about how the simulation experience might have offered insights or skills that they could apply in their future professional roles. There were other references to the future from participants as well.
Some suggested that the experience had helped them to understand something about their future work. For example, Marker 9 saw a connection between the scripts that they received in their marking bundle and how they might emphasise a particular approach to writing in future teaching: “The experience really drove home for me how important it is to support students to understand that they are always writing for a reader.” Marker 11 said, “I think, for me, the importance of explaining the task to the students. Because, so many of the ones I was reading just didn’t seem to answer the question.” These insights appear to have less to do with the activity of marking, and more to do with the opportunity participants had in the simulation to read a wide range of student work.

Other participants saw the rubric as potentially useful in future teaching. This is somewhat evident from Marker 3’s comments, reported in the section above on the Rubrics sub-theme, about how the experience had impressed strongly on them the importance of a good rubric in supporting making judgments as well as providing a starting point for feedback. Marker 8 and Marker 2 were even more explicit, suggesting that parts of the rubric might be co-opted for use in their future classrooms:

Marker 8: And we were also saying that back page is really good, in general, for every kind of writing assessment. It’ll be great to use that.

Researcher: The language conventions?

Marker 2: Yes. That was quite interesting. How it was broken down. It was something that we could use pretty much across the board.

Marker 8: Yes, you could just kind of manipulate the front [page] ones depending on what you’re looking at.

The idea of using rubrics from standardised assessments in the classroom is not new, but can be considered an unintended use of external assessments. What was more surprising to the researcher at the time was the possibility that the rubric from the simulation might be used in these participants’ future classrooms. As it eventuated, and is reported in Chapter 7, two of the seven participants who were interviewed in the follow-up phase used the rubric in their teaching practice. However, they did not use it directly to assess students, but when planning lessons, as an aide-memoire for features of persuasive writing. The use of the rubric outside of the marking simulation, in any capacity, was an unintended consequence of the simulation. In hindsight, though, if these preservice teachers felt comfortable with the rubric and confident in applying it because of the training that was an essential part of the simulation, then perhaps such a consequence ought to have been anticipated.

The general characterisation that can be drawn from this sub-theme is that there were aspects of the simulation that participants anticipated would be useful in their future careers.
Some found the rubric informative or inspiring, others found points of emphasis for their future teaching of writing. Though relatively few in number, these reported experiences and ideas suggest that at least some of the participants had learned something through the simulation that they anticipated would be useful for their future professional work.

**Exemplification of standards.** In both small-group interviews, there was some discussion about the exemplars that were used to illustrate the described features of each criterion in the marking guide. During these interchanges, a common theme was that exemplifying every score in the rubric was a practice that the participants hadn’t encountered before in their own education or on practicum. For example, Marker 1 said, “You gave us examples of lots of different types of writing. Whereas, usually when you’re in a school, you’ll get just good exemplars. You won’t get bad exemplars.” In the other small-group interview, there was also discussion around exemplification, and Marker 2 made a similar, but more general point, about how often thorough exemplification happens:

> Just to be able to look at, well, these all cover X, Y, Z but they’re all so different. That hardly ever happens. You’ll get one that’s an A. That’s what you’ll look at. To be able to see these four, they’re in the same range but you can see they’re succeeding in different ways, it’s so useful.

In general, it seemed that no participant was familiar with having exemplars of every score point in a rubric. The general practice they reported was being given a few exemplars to illustrate the best possible responses to an assessment.

Both groups also spoke about their personal experiences with exemplars in their university education, and it appeared that only having the best responses exemplified was a general practice. Another practice that both groups mentioned was being presented with exemplars that did not only exemplify the best responses, but where the exemplars illustrated a particular aspect that had been done well *without* being told which aspect was done well. For example, Marker 4 reported,

> I feel that with some education units, as well, where they’ve been, “Here are some examples. But we’re not saying which ones are good and which ones aren’t. So, you might just want to look at the structure and that sort of thing.” Rather than being, “this is a good example.”

Other participants generally agreed that the practice of receiving a number of examples of varying quality but no guidance about what parts of the marking guide they represented was common in a number of learning areas.

So, exemplification of all score points to support reliability in judgment was a new practice that the participants had not, apparently, encountered before in their education or teaching experience. In this regard, it may be said that the simulation provided insight for the
participants into new assessment practices. Though, whether the particular practice of working with an exemplified marking guide is able to be applied anywhere other than in large-scale marking might be considered. Nonetheless, the simulation appeared to have introduced at least this small amount of new knowledge and a new marking practice to these preservice teachers.

**Unique sub-themes.** As reported above, there were two sub-themes unique to the first small-group interview: Objectivity and Perceptions of large-scale marking. It also so happens that both of these are sub-themes of Outcomes and Achieved goals, and that one of the references was coded to both sub-themes. It was a statement from Marker 5:

I kind of wanted to get in and learn about these, kind of, mass-style marking environments and examples. And the biggest thing I think I took out of today, was this style of marking is not about assessing whether or not someone is good at, or bad at, something. It’s about seeing whether or not it’s evident in their work … And I think the penny dropped for me when we were looking through something, I think it was spelling or whatever, and you can’t mark someone positively for spelling challenging words if they’re not there. It doesn’t mean that they’re good at it or bad at it; it’s just that there’s nothing there to go on. And I think for me, and I can’t really articulate it, but there was a shift in my perspective once I’d learnt that. And, maybe, I got a little bit more respect for the whole standardised test thing, because I was like, oh I don’t know if I like this business. That’s what I kind of got from that today.

This response was interpreted as reflecting two ideas. The first was that Marker 5 had a different opinion of large-scale assessment after the simulation than before it. The change, according to Marker 5, came from gaining a different understanding of how student work is judged. That is, it is judged by markers who have only the student’s response, and no other knowledge of the student, upon which to judge the student’s ability. Judgments made only on the evidence found in a performance are, arguably, more objective than those made with additional knowledge about a student and their past work. In this regard, Marker 5’s response also reflects a new understanding of objectivity in judging student work. Other participants commented on objectivity as well. For example, Marker 1 said,

I also liked the advice you gave at the beginning, about how you had to separate, like, the purpose of marking like this is to separate, your feelings and emotions and stuff like that. That was really good. Because I think a lot of teachers who are new, maybe empathise too much with kids. And that can be a detriment in marking and also with behaviour management but mostly marking.

In response to which, Marker 20 said,

But I would say, even on the flip-side, my mentor who’s been in teaching for many years, he wanted at least two of his students to get a B. So, at the end of the course, he was like, I’m going to re-mark all their work and see if maybe I was a little too harsh or there’s a question where maybe they could just jump up one mark here. Or, vice-versa, if they just did stupidly well in one test and it didn’t look right, he was like,
maybe I need to re-mark and go back. I think there was definitely a case where he didn’t have any B-grade students in his class, and he needed one.

While no participant in the group explicitly mentioned objectivity during or in response to this exchange, the participants were certainly discussing marker objectivity. It is important to note, too, that there is a suggestion that what some of the participants had witnessed on practicum is not necessarily ideal assessment practice. Marker 20’s anecdote, for example, demonstrates the complex interplay between objectivity, reliability and fairness that exist in practical classroom assessment – the mentor-teacher either had particular targets to meet with regards to his grade distribution, or believed that the abilities of his students were not fairly represented by their current grades. Ideally, the previous assessments would have been well-constructed and reliably marked so that post hoc review would be unnecessary; however, as experience teaches us all, reality is rarely ideal. This is significant for this research because it highlights, yet again, the variability of participants’ practicum experiences and reinforces that the participants in the course all had different experiences of and with assessment. This is likely to have contributed to the variation in responses to participation in the simulation, over and above the uniqueness of the individual participant, and ought to be borne in mind when generalising about what effect the simulation is reported to have had on participants.

**Difficulties.** One of the guiding questions for the interviews was intended to establish what the most difficult aspects of the simulation were for the participants. Within this theme, three sub-themes were coded. Each sub-theme had relatively few references, but they are all reported here as the results have implications for any future marking simulations following the same procedures as this one.

**Noise.** The communication between participants as they marked was also discussed in Chapter 5. The researcher did not intervene in the simulation to discourage discussion between participants, and this was a deliberate choice, but as noted in Chapter 5, this active discussion among markers is not found in authentic large-scale marking experiences. The participants had mixed feelings about the discussions. Marker 5 was among the participants who generally favoured the more collaborative approach to marking, reporting,

I remember asking around, are we supposed to do these in silence or what? And I was prepared to just kind of zone in and focus, but we all just fell into this, turning around and asking the person next to us, so there was chatter the whole rest of the afternoon. And I resigned myself to the chatter because it kept me awake. It did. It kept me engaged. And I totally agree with your point [referring to a point Marker 9 made about the moderation activity] that I was better at it if I talked my ideas out, or I was giving more of a fair judgment if I was allowed to have that sounding board. To lock myself in a room and do 30 on my own, I don’t know whether or not I’d be giving the
best result possible, from my own perspective, to the marking, if I didn’t have someone to work it through with. (Marker 5)

This is in contrast to other participants such as Marker 9, who wrote in an email, “The only negative, for me, was the noise level during marking which was quite distracting. I’d bring head phones/earplugs another time.”

There were few occasions in the interviews where the researcher contributed his views on the level of discussion during the marking. At one point in the second small-group interview, Marker 8 asked, “Were you happy that we were working collaboratively?” To which the researcher replied,

What I was trying to do was replicate a real-life situation and make it as a simulation. And it’s true that the markers that I’ve worked with in the past, work in complete silence. I didn’t want to insist on that, though, because the markers that I’ve worked with are getting paid $40 an hour and, so, they have an incentive to hit their marking targets. To be efficient. But they’re also expected to already be experts as markers.

And, we expect them to have a certain level of experience and be professional and be capable of sitting at a computer screen for about seven hours a day, marking. Which is fatiguing, as [Marker 16] was saying. So, yes. While it was different for me to see people chatting a bit more … But it’s different.

At this point, Marker 13 observed that

I think that’s the difference between mass marking and classroom teacher marking. Because the office that I’m in doing my internship, they’re very often … like, asking each other’s opinions while marking. That’s their standard go-to. They don’t ever, really, mark alone.

To which the researcher responded,

And it’s not the case that the markers that I’ve worked with don’t do that. It’s just that it’s less frequent. So, often, they’ll find one of those cusp-ish scripts and not be sure themselves, even though they’re very experienced and have marked a thousand already. They’ll still find one where they’re like, can you have a look at this? Is this a two or a three?

In reviewing the transcripts, it’s obvious that the researcher was inconsistent in these two replies: he has both asserted that there is silence, and that markers seek help from each other when they find a script that they are unsure of. This inconsistency is left here to be judged, but in the researcher’s defence, the explanation is that experienced markers encounter scripts about which they are uncertain infrequently, and so the vast majority of time marking is done in near silence.

The researcher’s contributions to the conversation are included here not to be self-indulgent. Later in this chapter, the noise level caused by the discussions in the room is discussed in terms of its implications for any future marking simulations, and the researcher’s choice to not intervene is examined.
In all, there appeared to be a fairly even split between participants who reported the discussion between markers as being helpful and those who found it distracting. It is worth noting for any future marking simulations that some markers found the collaborative marking helpful and rewarding, while others would prefer not to mark in this way. This is a variable to be considered in planning any future simulations, and also in facilitating them. On the whole, though, it did not seem very detrimental to the experience. This was noted in Chapter 5, where the marking results were examined, but is also supported by the responses gathered from participants post-simulation: more often than not, noise was seen as a distraction but not a serious problem.

**Duration.** The second issue raised by participants as being problematic was the duration of the simulation. Marker 1 put it succinctly, saying, “The training, I understand why that had to go for that amount of time. I totally get that. Yes, it’s just the marking. I feel like it could be shorter and still achieve its goals.” A few others in that small-group interview commented on the difficulty of focusing on marking for more than an hour. One participant in the second small-group interview commented that 50 scripts was probably not necessary.

The simulation was already much shorter than a typical day of marking, in the researcher’s experience, and so it is interesting to consider whether replicating the duration of the authentic experience is necessary for the simulation. One of the hypothesised sources of benefit in the design of the simulation was exposure to a range of student work (Gambell & Hunter, 2004; Goldberg & Roswell, 2000; Masters & Forster, 2000), so it was quite surprising to the researcher that some participants did not see marking as equally important as, or more important than the training on the rubric. Changing the condition would lessen the fidelity of the reproduction, but if the learning outcomes could be achieved in a shorter period of time, then this might be considered for any future iterations of a marking simulation in the same vein as this one. It is, again, a variable to be considered in planning and facilitating future simulations. On the whole, though, there was only a handful of participants who found the duration to be too great and, based on the statistical analysis of reliability reported in Chapter 5, it did not appear to be detrimental to reliability (which would be expected if markers were unable to focus), even if some participants felt it was longer than necessary.

**Uncertainty.** As noted in other parts of this research, a seemingly common negative emotion associated with marking is uncertainty in making judgments. Many participants in this research reported experiencing anxiety, both in the phase one survey discussed in Chapter 4 and in the pre-simulation interviews discussed in this chapter. It is perhaps unsurprising, then, that in the post-simulation interviews, too, a few participants commented
on their feelings of uncertainty during the simulation. Marker 20 reported, “I spent a lot of
time in the morning doubting whether I was being consistent. Or whether my opinions were
accurate.” Others too, spoke of wanting to go back to previous scripts to change marks after
having read another similar one at a later point.

On the whole, though, there were only four participants represented in the references
coded to this node, so while uncertainty is a theme through different phases of this research,
and thus ought to be thoroughly considered, the strength of this negative emotion does not
appear to have been detrimental to the participants’ enjoyment of the day; nor, more
importantly, did it give cause for concern about participants’ wellbeing.

**Future participation.** The theme labelled ‘Future participation’ groups together two
sub-themes. In the interviews, the researcher asked whether participants would recommend
participating to others, and there was general consensus that they would. Beyond that though,
there was discussion about whether experiences like this could be integrated into ITE courses,
and in one of the small-group interviews, two participants said that not only would they
recommend it to others, but that they would be willing to participate in another simulation if
it focused on marking narrative writing (the simulation they had just participated in having
focused on persuasive writing).

The comments about participating in another simulation with narrative writing do not
warrant much discussion because they were comments made in passing by only two
participants: Marker 2 and Marker 12. In essence, these two participants suggested that
narrative writing is different to mark, so they would like an opportunity to train on a narrative
rubric and read students’ stories. Marker 2 said, “I actually would find the creative one a lot
more valuable. I think that’s intrinsically harder to mark. For me, anyway. I find I struggle
with stuff like that.” This reinforces the findings from the pre-simulation interviews that
many participants were motivated by experience and confidence: these preservice teachers
wanted experience in marking students’ narratives in order to be more confident in making
judgments about them.

The other sub-theme, though, is discussed in more detail, because of the participants’
passion about it.

**Marking experience as part of ITE.** If the researcher were to be asked about the post-
simulation interviews years hence, there is one quotation that he would certainly remember.
In many ways it is a blunt summary of one of the main justifications for this research:

You can qualify as a teacher and have marked nothing. That’s fucking ridiculous. A
lot of people have said that their first prac they marked nothing, a lot of people will
mark nothing in their second prac, and we do this theoretical assessment module where we do a lesson plan and we talk about how we're going to do formative assessment. No-one's marked anything. (Marker 12)

Marker 12 was quite animated – and evidently frustrated – in relating this point in the interview. It was clear that Marker 12 felt the lack of a requirement to have marking experience before qualifying to be a teacher was a major failing of ITE. The idea that it should be a course requirement was in fact brought up by Marker 16, who suggested shortly before Marker 12 made their statement that it had been the topic of a conversation prior to the group interview: “I think we mentioned it before, that we think this should be part of our course.” In response to Marker 12’s statement, Marker 8 volunteered that “I think it’s a specific problem in English where the marking is so, at times, really subjective … I think definitely for English teachers or any sort of humanity subject areas, it should be a requirement.”

There were similar comments made by participants in the other small-group interview, too. Marker 5, particularly, expressed disappointment in the assessment preparation available, and the potential future consequences of insufficient marking experience:

I was really surprised, coming into this course, and I don’t know whether you guys feel the same way, but I thought Assessment and Reporting, the unit, was going to be like, how to assess. Not the concepts of assessment for, as and of, learning. Do you know what I mean? And,…. I wasn’t expecting it to be covered in depth but, I thought we would have a class that would be like this. I thought it would be part of our degree. That we’d come in and we’d talk about pulling apart a rubric and working through how to assess people in quite a subject basis. I thought this would be part of my English course work and part of my Drama course work, and it’s pretty evident it’s not. It’s just kind of like, here’s your rubric. Good luck. You can find some more of these online, if you need some help. And, that feels a bit brutal for student teachers. And, though it comes with experience, what happens to the five-years’ worth of marking when you don’t have that experience? Where does it go? (Marker 5)

This participant was quite animated too, and the researcher detected concern and some frustration. So, both Marker 8 and Marker 5, in different group interviews, emphasised their bewilderment that marking itself was not something that they would have to do in order to become a teacher. No other participant offered a contrary point of view and there was a general level of approbation from others in the interviews, but it is difficult to estimate the extent to which others agreed with this sentiment.

The responses above provide substance for a few reflections about this research. The first is that these participants clearly found something valuable in the simulation that they felt was lacking from their teaching preparation up to that point. Bear in mind, as well, that these
participants were all in the final year of their education courses. These were not first-year students who would go on to learn more about assessment through their course work; they were approximately six months away from searching for jobs as newly qualified teachers. The other reflection is that these statements demonstrate the impact of this research. Up until the post-simulation interviews, the researcher had hypothesised only that a marking simulation would benefit the participants. These participants, though, clearly stated that not only did they think it was a worthwhile exercise for them, but that they felt marking experience was very much missing from their teaching preparation and that an experience like this one would be very worthwhile for every preservice teacher.

**Feedback for future iterations.** Finally, participants in both group interviews suggested what they would like to see if a similar simulation was run again in the future. The first was to do with the noise level, which some found distracting. The second, though, was that more time should be spent doing ‘team marking’ or ‘group assessment.’ Just after lunch, there was a session intended to be a moderation session. First, the group read several scripts together. Then, criterion-by-criterion, each participant selected a coloured card with a number on it and held it up to show what they scored the script on that criterion. The facilitator then asked individual markers to explain their mark. The participants reported that this was one of the most enjoyable and interesting parts of the day, and several suggested that more time be spent on it in any future iteration. In the previous chapter, it was noted that this session was originally planned to take place earlier in the day. So, if future iterations of this simulation were to be organised, more attention ought to be paid to scheduling, and additional moderation sessions ought to be considered.

**Summary of Themes in Post-Simulation Interviews and Emails**

There were four main themes identified in the small-group interviews and the email correspondence from participants collected after the simulation. The themes were Outcomes and Achieved goals; Difficulties; Future participation; and Feedback for future iterations. Several sub-themes were also identified, and these were grouped under a relevant theme. This section, however, generalises from these sub-themes and themes to summarise how participants perceived the simulation. This process is part of the Interactive Model of qualitative data analysis (Miles et al., 2014), because it is data condensation, one of the four process in the model. Each theme is summarised briefly, then a synthesis of participants’ perceptions of the simulation is given.
Some of the outcomes and achieved goals that participants reflected on post-simulation were anticipated by the participants in their pre-simulation goals and aims. Namely, many of the participants chose to volunteer for the simulation to gain experience and confidence in making judgments about student work, and in the post-simulation data, participants noted that they had gained both, thus meeting their expectations and aims for the day. The unanticipated ‘takeaways’ from the simulation, from the participants’ perspective, were to do with the processes on the day that were new to them: working with a very prescriptive rubric, and particularly the exemplification of each score point in the rubric through a carefully selected and curated selection of responses. These aspects of the simulation weren’t anticipated by the participants largely because they were unaware of them, not having observed or having been taught about them. On the whole, the participants reported having learned more about using rubrics and gaining confidence and experience in making judgments about student work.

The most difficult parts of the simulation for participants were focusing for extended periods of time and feelings of uncertainty about correctly scoring a script. These two aspects are related, because as the uncertainty led to participants discussing the scripts in front of them with their peers and other participants close by, there was a level of noise in the background (though it seems all of the noise was ‘on topic’ discussion of the pieces of writing in front of the markers). A handful of participants also suggested that the duration of the marking was longer than necessary, and that the same objectives could have been achieved with less actual marking. This suggests that for these participants, the parts of the day they felt to be the most beneficial were the training and moderation sessions, rather than the marking itself. Both the training and the moderation discussion relied on exemplars of the marking criteria, and it has been observed that exemplars are useful artefacts for these discussions and the learning that results (Carless, Kennedy, To, Lo, & Barrett, 2018; Wyatt-Smith & Adie, 2019). This point needs further consideration, as it was hypothesised earlier in this research that much of the learning from the simulation would stem from participants being exposed to a wide range of authentic student work through reading many pieces and having to make judgments about them. Of course, this may have still been a benefit, but the participants did not think it was as important as the exemplification of the rubric and the discussions of writing quality.

Participants were very positive about the simulation and there was general consensus that not only would they recommend others volunteer for a similar simulation, but that marking exercises like this ought to be a requirement of ITE. Two participants in particular,
in separate small-group interviews, strongly expressed their bemusement that practical marking was not a part of their courses. In addition to the general agreement among participants with the frank statements on that issue, a couple of participants also said that if the same simulation were to be run again, but with narrative writing as the focus, they would volunteer again. This implies that the participants saw the simulation as a valuable and worthwhile experience.

The participants also gave some suggestions for any future simulations. The first was that the discussion between participants during marking be managed to ensure an environment conducive to concentration. The second was that more time be spent on ‘group marking,’ or moderation, involving all participants. These suggestions are certainly worthy of consideration should any future iterations of this simulation be planned and delivered as a learning experience.

To summarise, the post-simulation data shows that participants found the experience, on the whole, valuable and that it provided them with the additional confidence and experience they were looking for when they volunteered. In addition to that, some participants felt strongly that their ITE courses lacked practical experiences like this simulation, and that they would recommend experiences like this be offered much more widely to preservice teachers.

**Discussion**

The purpose for gathering data from the preservice teachers who volunteered for the simulated marking experience was to provide evidence to answer the fifth and sixth guiding research questions introduced in Chapter 1: “What do participants in a simulation of large-scale marking report about the experience?” and “Is the experience reported by participants in the simulation of large-scale marking similar to the experience reported by experienced teachers who have participated in authentic marking operations?”

In the first section of the discussion that follows, the general summaries of the pre- and post-simulation data are compared and examined in order to synthesise concise, but precise, answers to these two questions.

**What participants report about the simulated marking experience**

The pre-simulation data analysis yielded quite clear findings about what drove participants to volunteer, and their motivation was closely connected, if not identical with, their goals and aim for the experience. Participants reported not feeling as though they had
enough marking experience, and so their goals were to gain more experience and more confidence in marking and making judgments of student work. These findings are in accordance with Grainger and Adie’s (2014) assertion that due to several factors, including a lack of published research into preservice teachers’ development as assessors and into how assessment is taught in Australian universities: “Teacher preparation courses graduate novice assessors, generally untrained in the important routine teacher tasks related to assessment” (Grainger & Adie, 2014, p. 97). Indeed, the two most animated responses from participants in the post-simulation interviews were on this exact topic, so it is clear that at least some participants were deeply concerned about their lack of preparation in assessment, and the general consensus in these data was that the participants wanted more marking experience before entering the classroom.

The analysis of the post-simulation data was similarly clear in its findings. The participants were positive about the experience, saying they would recommend participation to peers and colleagues, and several also expressed an interest in volunteering again if a similar simulation were to be run on a different genre of writing. The participants who spoke about the goals they had in taking part believed that they were now more confident and had more marking experience than they started with.

In addition to the general positivity, there were some comments about how to improve the simulation in future. A few participants suggested there could be less marking, and others commented they would like more moderation sessions, or ‘group marking’ as they referred to it. To turn to Grainger and Adie (2014) again, social moderation was reported in that study as likely to be effective in enhancing assessment knowledge and student confidence in marking, and so in any future iterations of the simulation, it would be worth considering a greater focus on that aspect.

So, the answer to “What do participants in a simulation of large-scale marking report about the experience?” can be summarised concisely as follows: Participants in this simulation of large-scale marking found the experience to be valuable as it provided additional marking experience and confidence that they had not acquired through their previous work experience or university courses.

This answer seems in some respects too simple. Questions remain about whether the participants would have felt similarly positive about any additional assessment practice, regardless of quality. In the absence of similar interventions, perhaps any experience is better than none. This problematises what ought to be a key finding of this research: from the participants’ perspective, the simulation was found to be effective and beneficial immediately
after the simulation, but what were the participants comparing it to? In the interviews, participants made comparisons between the simulation and practicum experiences, and between the simulation and their coursework. Whether these comparisons provide a sufficient frame of reference to judge the efficacy of the simulation is unclear.

Participants’ reports and teachers’ experiences of large-scale marking

The sixth guiding research question of this research asks to what extent the simulation experience is similar to what teachers report about authentic marking experiences. Chapter 4 provided the first part of the answer, presenting the results of phase one: a survey of teachers, most of whom had experience of large-scale marking. That chapter concluded that the findings of that survey were broadly similar to those of international literature looking at the benefits of marking. In particular, there was a variety of reported benefits for teachers that generally aligned with four categories identified by Falk and Ort (1998): clarifying their goals and expectations; deepening their discipline knowledge; learning more about students and their work; and developing insights that support their professional practice. In addition, several other benefits were identified in the survey responses, including maintaining links to the profession, networking and money. The present chapter showed that preservice teachers found the main benefits of participating in the simulation to be increased confidence and competence in marking, both as a result of gaining experience. While these outcomes appear on the surface to be different from those reported by experienced teachers in the survey and found in the literature, there may actually a reasonable amount of overlap between categories of benefit.

In the first place, survey respondents’ reported reasons for becoming markers included themes and sub-themes relevant to the simulation participants’ reasons for volunteering for the simulation. For example, one theme in the survey responses was ‘Gain experience,’ and the sub-themes grouped under ‘Professional development’ included ‘Increase skill or competence’ and ‘Increase confidence’ as reasons for marking. Thus, the two strong themes of experience and confidence that came through in the preservice teachers’ reported reasons for volunteering for the simulation are reflected in the reasons Australian teachers gave for becoming markers.

Additionally, there is a particularly obvious similarity in one of the closed-response survey items. When survey respondents were asked to identify statements about the effects of large-scale marking with which they agreed, 40 of the 43 respondents selected “increased your confidence in your ability to judge the quality of student work.” So, in respect of this
SIMULATED LARGE-SCALE MARKING

effect of participation, there is a close similarity between the survey respondents and the simulation participants.

Further, Chapter 4 presented some of the advice that survey respondents would give to colleagues about participating in large-scale marking. A large category of ‘Professional development benefits’ was identified, which comprised mostly vague allusions to such benefits but included a few minor themes that mesh with the idea of gaining confidence and experience: ‘Increase marking speed;’ ‘Increase confidence;’ and ‘Improve accuracy.’ It is far from clear that this is sufficient evidence to say that the simulation experience yields the same benefits as the authentic experience; it may, however, be enough evidence to propose that the survey respondents who alluded to ‘Professional development’ – who were in the majority – had confidence and experience in mind when they did so. This is especially so, if these answers are taken in concert with the reasons for marking that respondents gave in response to an earlier question in the survey.

But what of the marking itself: did the simulation participants’ reports of the marking experience reflect those of the survey respondents? The survey respondents reported several negative expectations of marking, including that marking would be time consuming, difficult, challenging and monotonous, in addition to more vague or broad negative emotions such as stress and feelings of inadequacy or incompetence. In this respect, the simulation participants’ experience deviates markedly from the survey respondents. The simulation participants did not report these as expectations of marking, nor did they suggest in the post-simulation interviews that they had felt stressed or struggled with the marking. The drawbacks suggested in Chapter 4 were not found in the results of this chapter.

So, does the experience reported by simulation participants mirror the reports of the survey respondents – experienced teachers – who had participated in large-scale marking? It is clear that the simulation participants’ experience is not at odds with the reasons for marking held by the survey respondents, nor do the participants’ reports suggest an experience contrary to that implied the survey respondents’ advice. So, there is some justification for asserting that the findings outlined in Chapter 4 and in this chapter are complementary and broadly aligned, even though no firm conclusion can be drawn about the closeness of the similarity.

Where the two experiences differ, though, was in the negative aspects of marking. The time-consuming nature of being on a marking panel while teaching full time, and the pressure and stress of scoring high-stakes and large-scale assessments where there are quotas and time pressures, were not replicated in the simulation. This reduction in potential negative
impacts was intended, as discussed in Chapter 1. There was an analogy drawn there of learning CPR on a resuscitation dummy, where the risk and negative consequences are removed so that learning can take place in a safe environment. When viewed in this light, this intended deviation from the authentic marking condition was then, in fact, a successful outcome of this research. The results reported in this chapter suggest that the simulation was able to provide a similar experience to the authentic large-scale marking, from which participants gained confidence, competence and experience in marking without having to feel the stress and pressure of the authentic experience.

Chapter Summary

This chapter presented the results of a qualitative analysis of participants’ expectations prior to the simulation and their impressions soon afterwards. It found that participants generally felt that the simulation was beneficial in that they perceived that it increased their experience and confidence in marking. The results help answer the fifth guiding research question, establishing that the participants valued the marking simulation experience. The results also help answer the sixth guiding research question: the simulation experience reported by participants generally accorded with teachers’ experiences reported in phase one of this research and in the literature on teachers’ experiences of marking more broadly.

Chapter 7 reports on follow-up interviews with simulation participants after a period of almost 12 months, and aims to ascertain whether the participants’ perceptions reported in this chapter changed after they begin their professional careers.
Chapter 7 – Phase Three: Follow-up

The follow-up phase aimed to investigate whether the simulation in phase two had any lasting impact on the participants. In particular, it was designed to determine whether participants’ perceptions of the experience changed over the intervening period of approximately 12 months, after they had begun their teaching careers. It was recognised that classroom experience may have led to participants re-evaluating the experience. The impact of teaching experience on assessment knowledge has been investigated to some extent in the literature and it appears, perhaps unsurprisingly, that teachers are likely to be more assessment literate than preservice teachers (Mertler, 2003). Since graduating, the teachers interviewed in this phase of the research had done more and different types of assessment as part of their daily work. By collecting participants’ retrospective evaluations of the experience, the researcher hoped to examine whether their increased assessment literacy changed their perceptions of the simulation’s value from how they reported it immediately afterwards. The aim of this chapter is to begin to formulate an answer to the seventh guiding question for this research: “After having transitioned from preservice to early career, will the participants’ impressions of the experience have changed compared to those captured immediately after the marking session?”

Of course, in order to make sense of this question, one must have in mind what the impressions captured immediately after the marking session were. These were presented in detail in Chapter 6, but a short summary of them here will frame the results that are reported and discussed in this chapter.

After the simulation, participants reported that they had achieved the goals they had when they volunteered to participate: to increase their confidence in marking and to gain experience. Several were emphatically positive about the value of the experience and a few strongly recommended that the experience be part of the ITE course they had completed. On the whole, responses were positive and indicated perceived value. Of central interest in this chapter, then, is whether the participants’ responses had become less positive, or perhaps more nuanced or equivocal, about the perceived benefits of the simulation.
Methods

Participants

After the simulation in August 2018, 15 of the 22 participants provided alternative email addresses where they would be contactable after graduation. These 15 participants were sent an email in the first week of June 2019, approximately six months after they completed their university courses, as it was intended that they would have been teaching for one semester by the time of the follow-up interviews. The interviews were planned to occur in the July school holidays. The email included an information letter and consent form that had been approved by the ECU HREC. A second reminder email was sent approximately two weeks later to participants who had not responded to the first email.

In addition to the recruitment procedures described above, the researcher’s primary supervisor was in personal contact with a simulation participant who had not provided an alternative email address. That participant was invited to contact the researcher if interested in being interviewed. After being contacted via email, the researcher forwarded to that participant the same information letter and consent form provided to all other participants in this phase of the research.

Taken together, the researcher successfully contacted eight participants: seven who were working as teachers, and one who had not yet completed the requirements of the degree and had taken up employment unrelated to education. Of the seven participants, two had completed a Bachelor of Education, and five had completed a Graduate Diploma of Education. Five were English majors with minors in Humanities and Social Sciences (HASS), one was a HASS major with an English minor, and the last was a Home Economics major with a minor in English. Whether the interviewees had found full-time, part-time or casual work was noted, but it was not used as a criterion to screen participants.

Sex is not hypothesised to be a relevant variable in this research. For anonymity and because it is not considered relevant, as in previous chapters the researcher has avoided using gendered pronouns for participants. The researcher acknowledges that in some unfortunate instances this has led to awkward expression, but he hopes the reader will understand the importance of anonymity for the participants.

Data Collection

For this phase, data were collected from one-on-one interviews. It was important that each participant saw the researcher as an interested peer, rather than as a marker-trainer, to
ensure the interviewees felt able to give open, frank and honest responses. To this end, the interviews were conducted at a time convenient to the interviewee and in a location chosen by them, though in the information letter the ECU Mount Lawley campus was offered as a possible location. Three interviewees chose to meet in person at the university campus, three more chose to be interviewed by phone, and one participant was interviewed via Skype. All interviews were conducted between July 8 and 21: the mid-year school holidays.

The interviews were semi-structured, using a set of guiding questions reproduced at Appendix I. The in-person and phone interviews were recorded using a smartpen and a digital voice recorder, and the researcher also made notes during the interview with the smartpen so that the notes were digitised and synchronised with the audio recorded by the pen. The Skype interview was recorded using the built-in recording functionality of the software application. That recording was downloaded immediately after the interview and then removed from the Skype cloud service where it was initially, automatically, stored.

**Data Analysis**

As with other phases in this research, coding was conducted only by the researcher, and was done in NVivo following the Interactive Model of qualitative data analysis (Miles et al, 2014) described in detail in Chapter 3. The interviews were deidentified, with names being replaced by the marker numbers used in the previous chapters of this thesis. So, for example, Marker 9’s comments in this chapter can be compared with comments attributed to Marker 9 in previous interviews, knowing that the same individual made those comments. This system is designed to help determine whether participants’ attitudes and perceptions changed over time or stayed the same. Given that a key aim for this phase of the research was to provide a longitudinal perspective on the reported experiences of the marking simulation, this ability to trace individuals across the research is important. Markers 6, 9, 15, 16, 18 and 21 were all interviewed prior to the marking simulation, while Marker 10 provided an email response prior to participating. So, there is some historical data for all the participants in this phase of the research, providing for longitudinal comparisons of their pre-simulation responses with the data collected in this phase of the research.

Markers 9 and 16 also participated in group interviews post-simulation, with Marker 9 participating in the first group and Marker 16 in the second. The other markers were not interviewed immediately post-simulation. This situation poses both a challenge and an opportunity. On the one hand, it is difficult to understand whether individuals’ attitudes to the simulation after the fact have changed over time, and a longitudinal analysis of individuals’
perceptions over time is not feasible. On the other, it provides an opportunity to hear more
detailed impressions from more of the participants. By comparing the impressions of a sub-
sample of participants almost a year after the simulation to those reported by a mostly
different sub-sample immediately post-simulation, a greater number of voices can be heard,
thus warranting more general statements about all participants’ impressions. This is
particularly the case where the sentiments expressed in the data collected at both time points
are in close agreement. So, while longitudinal conclusions for individuals was no longer
possible, the group’s perceptions across the three time points are able to be identified and
compared across almost a year.

Transcription for this phase of the research was approached differently from the
previous phase of the research, in that the recordings were not sent to a third party for full
transcription. Given the small number of interviews and the small number of guiding
questions in the interview schedule, the researcher elected to transcribe the portions of the
interviews relevant to the guiding questions himself. Then, the partial transcriptions along
with the full recordings were imported into a new NVivo project. In NVivo, audio files can
be listened to and transcriptions, annotations and codes can be linked to specific parts of the
recording during analysis. This functionality is useful when applying the Interactive Model of
qualitative data analysis (Miles et al., 2014), as the method involves transitioning between
processes where it is necessary; for example, returning to the data after drawing conclusions
to verify them, or further condensing data so they can be displayed in a way that helps in
drawing further conclusions.

Results

This section reports the themes and sub-themes that were identified in the interviews.
Initial analysis resulted in ten codes, or themes, being identified. During data condensation
(Miles et al., 2014), two of these themes were grouped under a similar but more general
theme that had been identified. The final list of codes is displayed below in Table 14, where
the indented themes are sub-themes of ‘Experiences of assessment.’ Each theme is presented
in detail in this section of the chapter, with a small amount of discussion where pertinent,
though the main discussion and summary of the results take place in the subsequent section.
SIMULATED LARGE-SCALE MARKING

Table 14 Codes ordered by number of references.

<table>
<thead>
<tr>
<th>Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiences of assessment</td>
</tr>
<tr>
<td>Assessment practices in different subject areas</td>
</tr>
<tr>
<td>Marking strategies or approaches</td>
</tr>
<tr>
<td>Initial Teacher Education</td>
</tr>
<tr>
<td>Relevance of simulation to professional practice</td>
</tr>
<tr>
<td>Outcomes</td>
</tr>
<tr>
<td>Future marking opportunities</td>
</tr>
<tr>
<td>Recommendations for in-service PD</td>
</tr>
<tr>
<td>Feedback for future iterations</td>
</tr>
<tr>
<td>Other positive aspects of the simulation</td>
</tr>
</tbody>
</table>

**Experiences of assessment**

In the course of the interviews, participants recounted anecdotes and experiences they had of assessment in the schools in which they now worked. Some, too, referred back to practicum teaching experiences they had after the simulation. Although there were 27 references coded to this theme, including the references coded to the two sub-themes, there were very few common ideas. On the whole it would be fair to say that each of the interviewees had different experiences of assessment, and that they were heavily influenced by context. For example, although Marker 18 had a steady supply of relief teaching, saying work was available everyday if they wanted it, this impacted their exposure to marking. Being a relief teacher meant that although Marker 18 had administered assessments and prepared students for assessments, they had not done any marking since beginning work as a teacher. This could be contrasted with the diverse assessment experiences of Marker 9, who had been teaching at a high school and also in two different universities, observing and participating in moderation practices and assessment activities in markedly different environments. The experiences of Marker 21 offered another contrast. This marker related that assessments were set at a departmental level in their school, and that all the assessments had been used before in previous years. Because of this ‘top-down’ approach to setting assessments, Marker 21 had not been given an opportunity to contribute to the assessment design, or even to be involved in revising older assessments before they were reused; Marker 21’s role as a teacher was to administer and score the assessments, rather than create them.
Even in the small sub-themes for this code, variability in experiences was evident. Markers 10, 15 and 21 had both taught in two subject areas: English and HASS. Marker 10’s impression was that “rubrics for HASS are much less specific than English,” while Marker 21 discussed a perceived difference that English was more skills-based and HASS more content-based, implying that the rubrics for English were more generic and the rubrics for HASS more specific. Marker 15 described the approach taken to assessment in the English department as “actually quite loose.” They contrasted that with the more regimented assessment schedule of the HASS department (similar to Marker 21’s experience described in the paragraph above), where the assessments were slightly more focused on “assessing content knowledge.” These few responses suggest that experiences of assessment vary not only between workplaces, but also among different subject areas in the same workplace.

In the other sub-theme of this code, ‘Marking strategies or approaches,’ there was again great variability. Three of the respondents, Markers 6, 10, and 16, referred to the way they organised their marking, noting that their colleagues had different approaches. Marker 6 spoke of applying one of the marking practices from the simulation by marking ‘blind’: the students submitted work electronically, without identifying information, then Marker 6 printed and scored the responses. This ensured objectivity and a focus on the response rather than on other knowledge of the student. Marker 10 spoke of experimenting with different strategies to find out what worked best, identifying that some teachers marked question by question, while others marked a whole paper at a time before moving on to another. The approach Marker 10 was taking at the time was to read through all papers with the rubric in mind and order the papers by quality prior to scoring, ostensibly to increase the reliability of the scores by ensuring increasing total scores would reflect increasing quality. Marker 16, however, reported scoring all papers on the first pass, then making a second pass to add comments and make any adjustments based on how the responses compared to the range of quality observed while scoring. All of these approaches are markedly different, and there are benefits and drawbacks to each one. In selecting and using different approaches, it may be that these markers were shaping their own professional practices.

So, across this theme and its sub-themes there are not many common threads other than that the experiences of each teacher were almost unique products of their particular teaching context. Perhaps this, in itself, is a commonality in a paradoxical way: the interviewees were now professionals responsible for their own practice, thus they were exercising their professional autonomy. Even Marker 21, who did not feel included in departmental assessment design, had the autonomy to implement informal and formative
assessment, even if the common assessment tasks that all classes sat were not within the scope of Marker 21’s responsibilities.

Later in this chapter, in the discussion section, the idea of assessment identity is revisited. To foreshadow that discussion, it bears saying here that the interviewees’ reflections on their experiences and choices in assessing the students in their charge may well be evidence of the development of these beginning teachers’ self-concept of themselves as professionals and as assessors. Given that some participants reported that the simulation was useful in learning about concepts such as reliability and objectivity, this may be the first suggestion that this research directly contributed to these beginning teachers’ PD.

**Initial Teacher Education**

As will be recalled from Chapter 6, one of the minor themes that emerged from group interviews with participants immediately after the simulation was whether such simulations should be included in ITE courses. This was expressed frankly and emphatically, in separate group interviews, by Markers 5 and 12 (who each participated in separate group interviews) that practical marking experiences ought to have been part of their ITE course. It was noted in Chapter 6 that although no participants offered a different perspective on the issue, and that there was a general sense of agreement, the extent to which other interviewees agreed could not be established. As noted earlier in this chapter, Markers 9 and 16 participated in post-simulation interviews, but the other participants in the follow-up phase did not. Five participants in follow-up interviewes referred to the idea that the simulation would be a worthwhile activity as part of the ITE course, three of whom were not present in the group interviews where the idea was raised immediately after the simulation.

The strength which with the interviewees advocated for the simulation to become part of the ITE courses they had studied varied, but there was a common, though not unanimous theme that the simulation ought to be included in teacher-training courses. For example, Marker 9 said, “I would recommend that it actually gets built into the training, into the teacher training at university, not just for English but especially for English.” And also, “I was really delighted to be offered it, but when I did it, I just thought it should be mandatory.”

Several markers articulated reasons for including the simulation in ITE courses. Marker 16 said that “doing it as a preservice teacher is really particularly good because there’s not a lot in your degree around that, you know, sort of assessment, and even things like rubric writing or creating assessments” Marker 21 similarly noted that the assessment content in the course was “all really theory and there’s not a lot of that practical application.”
Marker 6 said that in the “uni-delivered part” of their assessment education, the task set was to write a rubric, “and in fact, once you’ve written the question, it’s fairly straightforward to write the rubric from that.” Marker 21 saw the simulation as more important because it dealt with applying a rubric, not just creating one that may never be used. Marker 18 echoed the idea of the lack of practical content in the courses. They then disclosed that they had recommended in the University’s post-course surveys and feedback forms that the simulation, or similar experience, be included in the ITE course as a means of improving the course. Marker 18 also spoke about the possibility of the marking simulation becoming a “5 or 6-week program that is part of the course” rather than a one-off experience offered in addition to the course. In response to probing questions about what that might look like, Marker 18’s idea was that that the marker training “could have been done over a couple of sessions,” then participants would have consecutive weeks of practical marking sessions followed by “going back and talking about it.” This idea of an extended program of activities would be supported by, for example, the findings in Borko et al.’s review (2010) and Aubosson et al.’s (2009) work, which strongly suggested that the traditional one-off workshop approach of PD is less effective than teacher learning that can be sustained.

It is worth noting at this point that Marker 16, quoted in the previous paragraph, was also quoted in Chapter 6, speaking on the same theme in a post-simulation group interview. Marker 16’s opinion at that time was that the simulation ought to be part of the ITE course, and that there had been a conversation among a group of participants about it prior to the group interview. This opinion apparently went unchanged over the 11 months between simulation and follow-up interview, which may suggest something about the strength of Marker 16’s opinion.

Markers 10 and 15 did not suggest the simulation be incorporated into the ITE course. Despite this, both said that the simulation was valuable and that they would recommend it to preservice teachers. Marker 10 made only a short statement in favour of recommending the simulation for preservice teachers, but added, “I think it was valuable.” Marker 15 referred to the same deficiency as a number of other markers, saying, “depending on when your prac is, you don’t always get a lot of assessment during those periods, particularly for people doing the DipEd who only have only two relatively short pracs.” This pointed to a potential lack of experience or preparation in assessment during ITE. Despite noting this lack, though, Marker 15 did not suggest that the simulation be incorporated into the course, just that it would be useful for preservice teachers to do. So, while not being explicit about the simulation becoming part of teacher education, these participants still saw it as beneficial.
Marker 6 was the only Home Economics major (English minor) who participated in this research, and expressed an interesting perspective about the relevance of the persuasive writing marking, which was the focus of the simulation, across subject areas. Marker 6 said it was the explanation and exemplification of the task and rubric that was useful. Despite the task and rubric ostensibly being most relevant for teachers who would be teaching writing, they said,

"Yours was more literacy focused, so it was good, so everyone could do that, it shouldn’t just be English students. I think everyone could do that because I think it didn’t require a knowledge of English to do it. It just needed an understanding of sentence structure, or whether they were addressing the rubric. Really, you just needed to understand the rubric to be able to mark it. (Marker 6)"

They contrasted this with an experience in an assessment unit during their ITE course, in which students from several subject areas were enrolled:

"When we were in the other class, it was all trying to be subject-based, so you can’t compare to anyone else in the room. Whereas when we were with you, we could do that exercise where everyone did that exercise where “I reckon it’s this” or “I reckon it’s that” and you’ve got a yardstick to gauge yourself against. (Marker 6)"

So, for Marker 6 one of the sources of benefit of the marking experience was in being able to mark in a marking panel, regardless of the task’s relevance to a major subject area. Marker 6 apparently found the general approach to marking and the experience of comparing and moderating judgments useful and valuable independent of the subject content.

In all, there was a unanimous feeling that the simulation should be recommended to preservice teachers as a valuable learning experience. The surprising part for the researcher was that most interviewees went beyond simply recommending the simulation to others, to advocating that it become a mandatory part of the university courses they had completed approximately six months prior to their follow-up interview. One inference that can be drawn from this is that the participants’ perception of the simulation’s value did not appear to have weakened over time. As noted previously, only two of the participants were interviewed immediately post-simulation, but even without direct comparisons for most interviewees, this sentiment mirrors those reported by other participants in those post-simulation interviews. There was a reasonably strong positive sentiment about the experience in terms of its relevance for preservice teachers expressed in both this third phase and the second phase of the research.
Relevance of the simulation to professional practice

Four of the interviewees referred to using specific tools, processes or feedback from the simulation in their professional practice. Marker 18 related an anecdote about being given the task of preparing a class of students for a large-scale writing assessment. As part of preparing for the class, Marker 18 used the marking rubric from the marking simulation to guide them in identifying aspects of writing to emphasise to students. Marker 16, too, located the notes and marking rubric from the simulation when preparing to teach persuasive writing to a class. Marker 6 spoke of drawing on the marker training aspects of the day, and explaining a rubric to a class of senior students. They followed the same processes: displaying a criterion, talking through the qualitative differences in the categories, and asking the students to identify those qualities in sample responses.

Both Marker 9 and Marker 21 explicitly referenced an analogy used in the simulated marker training: the ‘bucket’ analogy that was used to explain why half marks are not assigned when applying rubrics in a large-scale marking operation. The analogy compares choosing a score in a criterion with sorting objects into one of a set number of buckets by choosing the bucket that contains the most similar objects. It is a means of explaining that selecting a described category in a criterion is a forced choice, and so the category descriptors must be carefully written so they can be used to justify difficult classifications. This analogy can be seen as one of the marking processes of the simulation that these two beginning teachers saw as having continuing usefulness in their professional work.

Markers 9 and 18 also said that they had included the marking simulation on their Curriculum Vitae (CV), with Marker 9 also including the agreement rates received in the feedback post-simulation. The researcher’s impression on hearing that these interviewees had included the simulation on their CVs was that they felt the experience slightly enhanced their desirability to potential employers. Potential employers are likely looking for employees with more experience, and the participants saw the simulation as giving them relevant experience in marking. Of course, this is largely speculation, but it appears to be a reasonable explanation for including the activity on a CV.

Here again, there is variety in the range of responses grouped together into a single theme. The unifying idea is the continuing relevance of the simulation in the teachers’ professional lives. Some had used the tools of the simulation, the rubric, others had used the processes, and still others had used the experience to frame their reflective practice. All of these things are, perhaps, small contributions in the scope of a teacher’s professional duties.
and responsibilities. The key aspect to note is that there are several ways in which the simulation experience provided some practical and relevant utility to the participants, even if they were minor.

Outcomes

The label for this code, ‘Outcomes,’ is perhaps a little ambiguous and so deserves some explanation. The references that were coded to this theme were those where an interviewee spoke generally about a benefit of the simulation experience, rather than specifically about an aspect of their professional practice in which they used a process or artefact from the simulation. The qualitative difference is in generality. An interviewee stating that the simulation made them confident (as some did) is a general benefit, whereas an interviewee saying that they used the rubric from the simulation in teaching their students about writing (as previously described) is a specific way the simulation has contributed to the teacher’s professional work. In this section, ‘outcome’ refers to a general benefit, rather than a specific instance of the simulation providing utility.

As discussed in Chapter 6, participants’ reports about what they gained from the marking simulation revealed a major theme: more confidence and more experience in marking. This theme was also present in the follow-up interviews conducted in phase three, even though most of the interviewees had not taken part in a post-simulation group interview. In the words of Marker 6, “What the marking simulation did for me was give me the confidence to mark because I was very concerned about being consistent or in line with anyone else.” And later, “Having done your simulation made me feel like I was confident in my marking.” Markers 9, 10 and 18 all referred to the simulation as being confidence building as well. Marker 16 gave a more mixed response, reporting feeling mostly confident when marking, but still relying on colleagues for feedback about marking and that sometimes they did not feel “super-confident” about marking. That said, Marker 16 also indicated feeling more confident having done the simulation than they would have otherwise.

Marker 15’s response was unique. The outcome that “stayed with me the most,” they reported, was seeing the range of quality of student work. They noted that up until that point, “I hadn’t had a lot of exposure except for in my prac, which were quite short, to what to expect from a Year 7 student versus a Year 8 versus a Year 9 versus a Year 10.” But shortly afterward they tempered this statement with, “it was kind of a one-off, one afternoon thing, so, I don’t think it had a massive impact, but it was definitely good to have that kind of
exposure.” In short, the simulation provided an experience of student work that was lacking up to that point, but that there is only so much that can be achieved in a short workshop.

In addition to referring to confidence, some of the interviewees named some other key aspects of the day they perceived to be useful outcomes. Marker 21 coined a phrase, stating that the “marks-centred approach” demonstrated in the simulation had shaped their own approach to assessment. Here they were referring to the way large-scale marking is objective and anonymous – the only evidence that a marker has on which to make a judgment is the student’s response and the rubric; the decision does not take into account any knowledge of the student because the marker has none. This outcome can also be considered in the light of the Marker 6 having identified ‘blind marking’ as a benefit, as discussed above. Marker 6’s reference was coded to ‘Experiences of assessment;’ however, because Marker 21 did not refer to using blind marking, and spoke more generally about objectivity, their response was coded under ‘Outcome.’ This is perhaps a fine distinction, but it is consistent with the definition of ‘outcomes’ given at the beginning of this section. In a similar way, Marker 9’s mention of the “importance of managing bias” was classified as an Outcome, as it implies that while their understanding of objectivity was gleaned or enhanced by the simulation, it was as a general principle rather than as a work practice.

Markers 6, 9, 10, 18 and 21 all referred, indirectly or directly, to marking reliability and agreement. Markers 9 and 21 mentioned moderation specifically as an outcome from the day, while the other markers spoke more about the discussions they had with colleagues while making judgments, or about being ‘correct’ or ‘accurate’ in their marking. Given most markers only referred to it indirectly, this outcome of marking reliability and agreement has largely been inferred by the researcher during the analysis. His impression was that these interviewees were aware of the importance of reliability and agreement in making judgments about students. This inference is supported by the fact that the concepts of reliability and agreement were emphasised in the marker training portion of the simulation, and while no causal link between that training and the interviewees’ conceptions about or approaches to marking can be established, it is worth noting the coincidence of these concepts that were embedded in the marker training also being referred to and discussed by the participants in these follow-up interviews.

On the whole, the outcomes – the general benefits – of the simulation as reported by participants in these follow-up interviews were that the simulation gave them confidence in their marking, experience of marking, and that it also provided or supported their
understanding of concepts like objectivity, reliability and agreement as they pertain to marking.

**Future marking opportunities**

A question asked of all interviewees was whether they would consider, now or in future, applying for large-scale marking. The intent of this question was to gauge whether, perhaps, the marking simulation could have encouraged them to pursue such opportunities. Often the question was preceded by inquiring whether the interviewee was aware of colleagues marking exams or national assessments in addition to their teaching work.

Marker 21 was the most certain about pursuing marking opportunities in future, saying, “Because you’re marking such a variety it’s helpful, so you don’t get stuck in your bubble. Helpful so you don’t start thinking, ‘well this is a top mark because it’s the best I’ve seen at this school.’” Marker 15 contributed a similar view, suggesting that marking senior school exams would provide insight into what the markers are looking for in exam responses. Both of these comments mesh closely with the reports of experienced teachers about the benefits of large-scale marking, which were presented in Chapter 4.

Markers 9 and 18 both said they would consider applying to be markers in future, but that they felt they would need more teaching experience first. Although, while Marker 9 said that experience would be necessary, Marker 18 had a different perspective:

I don’t think anyone would give me the job of being a marker right now just because in their minds I haven’t had enough experience in a classroom. But I don’t actually think you need extensive experience; I think you just need to understand what you’re marking and be able to interpret it and mark it.

This is in contrast with Marker 6, whose contract was short-term and who was anticipating having to search for work at the end of the year. They said, “I feel I can apply for those sorts of jobs as well, now, whereas perhaps I wouldn’t have had before.” It will be remembered from Chapter 1 that there is a perception that teachers need experience in order to apply for marking work. That some of these beginning teachers, who have been trained as large-scale markers and had feedback demonstrating their reliability as markers, still see teaching experience as a real or perceived barrier to taking up marking opportunities, is illuminating. The perception that experience is mandatory is, it seems, entrenched.

**Feedback for future iterations**

In the course of these semi-structured interviews, the researcher invited, when it seemed natural and appropriate, feedback on whether there was anything about the simulation
that they would like to see changed if it were to run again. In particular, he asked about the ratio of training to marking, but also whether the interviewees had any other suggestions for future iterations.

Markers 10, 16 and 21 indicated that they themselves would participate in another simulation if one was offered for narrative or another genre of writing, with Markers 10 and 21 suggesting that they would be interested in a HASS marking simulation. It will be recalled that in the post-simulation interviews reported in Chapter 6, a minor theme identified was that participants would be willing to participate in another marking simulation on narrative marking. Because the recommendation was straightforward, this theme was dealt with briefly in that chapter. However, as shown there, it was Markers 2 and 12 who spoke about the idea in the group interview, and neither of these two were interviewed in this phase of the research. This, therefore, is an example of an idea that, even though it has been expressed by different people, is likely to have persisted among the participants since the simulation.

Marker 9 gave feedback about the noise level in the marking room resulting from discussions between markers about the scripts in front of them. It may be remembered from Chapter 6 that Marker 9 provided similar feedback in an email, post-simulation. That this negative aspect of the simulation was still well-remembered and warranted further comment perhaps speaks to the strength of Marker 9’s feeling.

The interviewees were also asked about the ratio of training to marking, and whether they felt either was too long, or whether less time could have been spent on one or the other. In general, there was little response other than general agreement that both parts were necessary and in good proportion. However, Marker 18 did comment that

I could have spent longer sitting there doing the simulation, the practical stuff, myself. But then, like I said, I’m quite happy doing that kind of stuff. I enjoy that. So, someone else may not have liked just sitting there reading it. But I wanted to.

Marker 18 added that perhaps there could have been an option for markers who similarly enjoyed marking to carry on for longer if they wanted to. This is interesting because no marker finished the allocation of 50 scripts in the time allocated to marking, so had this been considered in the original design of the simulation, the ‘finish point’ of the simulation could have been set as the end of the allocation, rather than a point in time, or even set as the point in time where the participants felt ready to finish. In other words, the design could have allowed the participants a much more self-directed experience. Whether that would have improved the experience for most participants is unclear, but this could be examined if there are future iterations of this marking simulation.
Recommendations for in-service professional development

As well as asking interviewees whether they would recommend the marking simulation to preservice teachers, the researcher also asked whether it would be something they would recommend to their colleagues as in-service PD, if it were to be offered. All seven interviewees indicated that they would recommend it, but in the accompanying comments a minor theme emerged.

Markers 16, 18 and 21 said they would recommend it, but they all anticipated that their colleagues might be reluctant. Marker 18 said that experienced teachers might “think they know better,” and Marker 16 expressed a very similar sentiment, commenting that an experienced teacher might say, “oh, we know how to do that.” Marker 21 characterised this attitude as an “I know it all” one. These three interviewees were not speaking poorly of their colleagues and no slight was intended. They were simply anticipating that experienced teachers may not see the benefit in a marking simulation.

This perception surprised the researcher, because the results discussed in Chapter 4 show that some, probably many, experienced teachers do see clear benefits in marking. Additional support for the proposition that teachers see value in marking comes from PD opportunities such as the marking experiences offered by NESA that were referred to in Chapter 1 (NESA, 2019). It is unlikely that such opportunities would be offered if experienced teachers in NSW saw no benefit in marking experiences, and it is unlikely that the NSW and Western Australian contexts are so very different that there would be no interest from Western Australian teachers in similar, marking-focused in-service PD.

Additionally, there was an awareness from a number of the interviewees of colleagues having been markers for large-scale assessments, including the externally administered Year 12 exams. Marker 16, who was one of the interviewees to suggest that experienced teachers might be reluctant, spoke about colleagues marking the senior school exams: “There’s generally quite an awareness of [exam marking] and quite a lot of teachers that do do it, rather than it being something that they’re just not involved in.” Perhaps, though, these ideas are not as inconsistent as they might appear from first appearances. If Marker 16’s colleagues were experienced and did mark operationally, then their feeling of not needing to do a simulated marking exercise may be very well explained by the fact that they had authentic marking experiences to learn from.
Other positive aspects of the simulation

Throughout the interviews there were a number of positive comments that do not, unfortunately, share commonalities that enable them to be grouped together under a single code. However, the very positive sentiment that some interviewees had towards the simulation should, the researcher feels, be reported because they are useful in answering an overarching research question: essentially, whether simulated marking can offer a valuable learning experience for preservice teachers.

Marker 16 spoke of the marking simulation as way of providing preservice teachers who had had different practicum experiences a way to remediate missed opportunities to mark real student work:

At that stage of being a preservice teacher, I still think that looking at the student work was really valuable and more from a fact that depending on what you’d done on prac and what your mentor teacher had done, you might not have had access to as much student work or marking as others, you know you talk to different people and have had different experiences on ATP some might mark everything. My mentor teacher let me look at things and give them a mark, but I had nothing to do with what their actual mark was. She went back and marked.

Marker 18 had a broadly similar response, in that they focused on the authenticity of the marking and its capacity to make up for not having had an authentic marking experience in their ITE course, saying,

Sitting there and doing simulations on the computer, it was actual marking, it wasn’t just someone standing in front of a chalkboard telling me, “this is what you should be looking for.” It was actually me physically having to look for it and find it myself.

Marker 21, on the other hand, gave a general impression: “I thought it was super helpful. I think it was probably one of the most helpful things I did in the year, grad, and it wasn’t actually part of the degree.”

These positive comments about the experience are not sufficient to argue that the marking simulation was successful or viable, but they echo some of the key themes that have been presented in this section and in other parts of this research. The ideas were that practical marking experiences are largely absent from the ITE courses these participants completed and that marking experiences are also often inadequate in (or absent from) students’ practicum experiences.

Discussion

The aim of this phase of research was to collect evidence to help answer the seventh guiding question: “After having transitioned from preservice to early career, will the
participants’ impressions of the experience have changed compared to those captured immediately after the marking session?” One challenge to the data collection in this phase was that only two participants were interviewed both post-simulation and in this follow-up phase, meaning that direct comparisons over the almost-twelve-month period were scarce. It was found, though, that those two participants expressed quite similar opinions of the research in both interviews. Beyond that, it was found that the responses gathered in this phase broadly reflected the themes that emerged from the post-simulation interviews. So, while it is not possible to give a direct answer to the guiding question by examining how a range of individuals’ opinions changed over time, it is argued here that impressions of the marking simulation as a whole persisted over time. And, given the enthusiasm with which five of the interviewees advocated for the simulation (or a longer program of activities similarly based on simulated marking) to be included in ITE courses, the strength of their opinions had not diminished by much, if at all. In other words, it is clear from the results that the participants’ estimation of the simulation’s value had not significantly changed in the intervening period – they were as positive about the experience as their peers who were interviewed post-simulation, and they valued it enough to suggest that it ought to be part of ITE.

However, beyond the simulation being a generally positive experience that the participants valued, a question arises as to whether the experience actually supported the participants in developing their self-concept as assessors and the development of their assessment identity. As discussed in the literature review in Chapter 2, assessment identity is a relatively recent term that has been used to conceptualise teachers’ sense of being an assessor, and builds on a range of earlier work looking at identity and its connections with assessment (Looney et al., 2018; Xu & Brown, 2016). The interviewees in this phase of the research had begun their teaching careers, and it would be expected that they would be still developing their teacher identities and, probably, their identities as assessors and judges of student work.

As reported in the results section, the responses of three interviewees were coded under the minor theme of ‘Marking strategies or approaches’ described previously. The three teachers described their marking processes, and each was different. One marked ‘blind’ by anonymising student work to increase objectivity, one ranked papers before scoring in an attempt to ensure reliable scoring (even though the researcher does not believe this would be an effective strategy), and one scored all papers before revisiting each to write feedback and adjust scores where necessary. As stated previously, this diversity of approaches might
indicate that these teachers were thoughtfully and deliberately making choices about their assessment practices, which suggests they had formed an idea of themselves as assessors who, in the way professionals ought to, decide on how to practice their judgment in a way that is responsive to their contexts.

Beyond this, there were specific practices and tools from the simulation that several of the interviewees used in their classroom assessment practices. In their responses, the participants mentioned a wide range of practices and tools from the simulation that they found relevant to their practice. Markers 16 and 18 reported using the rubric from the simulation in their preparation to teach persuasive writing, while Marker 6 used the training practices from the day to help students understand a rubric and its descriptions of different levels of quality. Marker 9 and 21 reflected on an analogy from the training to remind themselves to make definitive categorical judgments rather than ‘hedging’ between described categories in a rubric by using half marks. While no single aspect of the simulation appears to have been useful for all participants, again it seems that participants were making choices about their teaching practice, and most had found useful tools from the experience that they incorporated into their teaching practice as they saw fit. It could be argued, then, that the simulation supported these participants’ professional practice to some degree.

Beyond the practical utility some participants derived from the experience, there was a more general theme of the simulation having built self-confidence in their marking ability, provided new approaches to assessment, and raised awareness of reliability and objectivity in assessment. These themes, particularly building self-confidence in assessment, also may have supported these preservice teachers in developing their assessment identity.

**Chapter Summary**

This chapter presented the results of seven interviews with participants almost 12 months post-simulation. It found that the interviewees’ responses about the simulation were consistent with those collected immediately post-simulation, despite there being few participants who were interviewed both post-simulation and in this follow-up phase. In particular, all recommended the simulation as a worthwhile activity for preservice and in-service teachers, with some going so far as to advocate for its inclusion in ITE courses. The answer to the seventh guiding research question was discussed, and it was found that the participants’ reports of the experience had not markedly changed since they transitioned from preservice teachers to beginning teachers. The chapter closed by considering the slightly tangential question of whether, based on the participants’ responses, the simulation provided
them with actual professional benefit, beyond their reported positivity and perception of value. It was found that aspects of the simulation had practical utility for some interviewees, that some were making deliberate and thoughtful decisions about how they assessed, which may have been influenced by the simulation. Also, most participants reported the experience was confidence-building, which might suggest that it was beneficial in building their self-concept.

In Chapter 8, these threads of thought, along with others from the earlier phases of the research, are picked up again. The answers to the guiding questions, which up to this point have been specific to each phase, are integrated and combined, leading to a proposed answer to the overall research question.
Chapter 8 – Synthesis and Discussion

The research project to this point has been considered largely as three separate and distinct phases, though some note has been made of where themes have persisted across phases. The aim of this chapter is to summarise all that has been learned in the three phases of this research and to draw a conclusion that offers the best possible answer to the research question:

To what extent can participation in a large-scale marking simulation provide, for preservice teachers, a professional development experience similar to authentic large-scale marking?

In the first chapter, seven guiding questions were put forward, each intended to contribute to the answer for the research question. The guiding questions informed the design and direction of each of the project’s phases. For ease of reference, the guiding questions for the research were:

1. What research evidence exists about the benefits of participation in large-scale marking?
2. What benefits do experienced Australian teachers report they have gained from participation in large-scale marking?
3. What are the opportunity costs, difficulties or other impacts reported by Australian teachers about their participation in large-scale marking?
4. Can large-scale marking be effectively simulated?
5. What do participants in a simulation of large-scale marking report about the experience?
6. Is the experience reported by participants in the simulation of large-scale marking similar to the experience reported by experienced teachers who have participated in authentic marking operations?
7. After having transitioned from preservice to early career, will the participants’ impressions of the experience have changed compared to those captured immediately after the marking session?

Each of the three phases of the research was relevant to different guiding questions. Table 15 summarises which guiding questions are relevant to which of the preceding chapters, and which phase of the research.
This chapter begins by reviewing the findings of each phase of the research in respect to the answers that phase offers to the relevant guiding questions. Next, the answers to the guiding questions are considered in terms of how they relate to one another, and to the research question. The chapter concludes with a synthesis of the findings of the three phases, to finally arrive at an answer to the research question.

**Summary of Results of Phase One**

The aim of phase one was to establish whether Australian teachers report similar experiences of participating in large-scale marking as teachers from other contexts as published in the literature (Falk & Ort, 1998; Gilmore, 2002; Goldberg, 2012; Goldberg & Roswell, 2000). The phase was shaped by the first three guiding research questions, and accordingly sought to establish: what is known about the benefits derived from large-scale marking; whether those benefits are also identified in the Australian context; and whether there were negative aspects to the experience that may not have been reported in the literature. This phase of the research can perhaps best be understood as setting a baseline or painting a backdrop for the next phases of the research.

The literature review in Chapter 2 contributed an answer to the first guiding question, finding that teachers’ experiences of marking and the claims of benefit were varied and dependent on context. However, four relatively broad categories of benefit for teachers were identified by Falk and Ort (1998): clarifying goals and expectations; deepening discipline knowledge; learning more about students and their work; and developing insights that support their professional practice. These categories cover a wide range of experiences reported in the relatively limited literature available.

The second and third guiding questions were addressed by an online survey of Australian teachers. In terms of the benefits of large-scale marking, the topic of the second question, the findings were that the responses from Australian teachers surveyed broadly aligned with Falk and Ort’s four categories of benefit. The survey results also indicated that
Australian teachers identified money as a relatively important incentive, along with networking with colleagues and maintaining links to the profession when on parental leave.

The third guiding question invited inquiry into the drawbacks or costs of marking, as well as barriers to entry. These more negative aspects of large-scale marking were not found to be much discussed in the literature, but the survey results highlighted stress, anxiety and time pressure as negative aspects of the marking experience. Time was also seen as an obstacle to participation, both by those who had marked and those who wished to but had not yet been able to. For regional and non-metropolitan teachers, distance was a barrier to taking up marking opportunities, which may reflect Australian geography. There was also a suggestion from surveyed teachers who had not yet been markers that teaching experience was a barrier to entry. Despite these negatives, though, there was broad agreement from teachers who had marked that the experience was to be recommended to colleagues.

Summary of Results of Phase Two

The central focus of this research project is the simulation of large-scale marking that was carried out in phase two, and the three guiding research questions relevant to the simulation were addressed in Chapters 5 and 6. Each of these chapters approached the simulation from a different perspective, though. Chapter 5 took the researcher’s perspective, which itself had aspects of two different viewpoints: the perspective of a ‘marker manager’ interested in marking rates, interrater reliability, and rater severity, and also the perspective of a ‘simulation facilitator’ who reflected on the authenticity of the experience compared to his past marking experiences. Chapter 6 sought the participants’ perspectives – their perceptions of the good and bad parts of the simulation and their judgments about whether there was value and learning in the experience.

The researcher’s perspective provided some grounds for believing that the marker training was effective enough to support acceptable levels of interrater reliability, and a Rasch analysis indicated that the preservice teachers as a group applied the marking guide to the scripts approximately as consistently as the ACER marking panel. This was taken as quantitative evidence of the simulation’s efficacy, though with many caveats regarding sample sizes and measurement errors. Additionally, the researcher reflected that he perceived the simulation to be authentic in most regards, with the exception of a few logistics issues (with room mix-ups and timings) and a choice to not discourage participants conversing during marking. The answer to the fourth guiding question, “Can large-scale marking be effectively simulated?,” is contingent on the meaning of ‘effective.’ If that is taken to mean
Effective training and an effective replication of an authentic marking experience, the answer is yes. The simulation appeared to have been successful in training the participants to mark reliably, and the researcher perceived it to be similar to his past experiences of large-scale marking.

From the participants’ perspectives, as gathered in the pre- and post-simulation interviews and emails, the simulation was effective as well. The participants reported volunteering because they were seeking confidence and experience, and after the simulation they reported feeling more confident and that they had gained experience in assessing student work. Additionally, there was positivity about the experience and a positive reaction when asked whether they would recommend it to others. So, the brief answer to guiding question five is that participants reported gaining experience and confidence, and that they valued taking part.

In the discussion of the participants’ reported experiences, comparisons were drawn with the benefits and drawbacks identified in phase one. This was done in order to address the sixth guiding question, which sought to uncover whether the reported experiences of authentic and simulated large-scale marking were similar. It was found that there were identifiable similarities in the reported experiences. For example, in one of the close-ended survey items, markers identified gaining confidence in judging student work, and for many participants this was both an explicit aim as well as an identified outcome of the simulation experience. Some negative emotions associated with marking, particularly fatigue and feeling uncertain about making correct judgments, were also reported by simulation participants, although to a much lesser extent than the markers surveyed in phase one. However, even this supports an answer for the sixth guiding question that the experiences were more alike than they were different.

The answers to the fifth and sixth questions provide further support for the answer to the fourth guiding research question, which is that a large-scale marking experience had been effectively simulated. The similarity of the participants’ reported benefits and those of markers who had experienced authentic marking, along with the quantitative data and the researcher’s reflections, all point towards a simulation in which participants achieved the goals and aims they had for participating and from which they derived increased confidence and experience – that is, an effective simulation.
Summary of Results of Phase Three

The follow-up phase of this research, phase three, was concerned principally with whether the perceptions and opinions offered by participants in phase two persisted over time, after participants had taken up employment as teachers.

One of the challenges in this phase of the research was panel attrition. The follow-up phase had been intended to provide a longitudinal perspective at the individual level, with interviews pre- and post-simulation in phase two, and then a follow-up interview in phase three. In the end, there were seven interviewees involved in the follow-up phase but data for all interviews was collected from only two. However, the other five phase-three interviewees had been interviewed or answered questions by email prior to the simulation, so there was some longitudinal data for all interviewees in this phase. Additionally, comparisons of each sample’s perceptions still afforded a longitudinal perspective at a group level.

In brief, the finding of this phase of the research was that perceptions of value persisted strongly. Some of the interviewees were adamant that the simulation was beneficial, and advocated for it to be included as part of ITE courses. All the participants were positive about the experience and reported that they would recommend it to preservice teacher and colleagues.

In addition to this finding, it was discovered that four interviewees had adopted sayings or doings (Kemmis et al., 2014) from the simulation for their own classroom practice, and that interviewees had increased confidence in their ability, benefiting from the exposure to a range of student work they received in the simulation, and adopting approaches to their marking work consistent with the principles of reliability and objectivity that were emphasised in the simulation.

On the whole, the findings of phase three were consistent with the perceptions collected immediately post-simulation, indicating that the perceived value and benefit lasted over a period of time that included a major career transition. Further, some direct impacts of the simulation on the participants’ professional work were identified, such as using the rubric or training techniques from the simulation.

Synthesis

The findings of each of the three phases of the research have been presented above. However, each phase is largely considered in isolation from the others, and the thesis presented here is that each phase contributes to the answer to a single question. In the
sections that follow, there is some repetition of the findings reported above. However, each section draws links between the findings from each phase, synthesising three main conclusions of the research: the benefits of large-scale marking reported in international research are largely generalisable to the Australian context; fidelity may not be a priority in simulations of large-scale marking, but it is important to consider; and that more practical marking experiences ought to be included in the ITE courses the participants completed. These conclusions are drawn from the answers to the guiding research questions, and while they themselves are not answers to the research question, they encapsulate the findings of the research. In the final section of this chapter, they form the basis of the answer to the research question.

The Benefits of Large-scale Marking for Teachers and Preservice Teachers

The aim of the first phase of this research was to establish whether Australian teachers report similar experiences of participation in large-scale marking as teachers from other contexts, which were identified in the literature review in Chapter 2 (Falk & Ort, 1998; Gilmore, 2002; Goldberg, 2012; Goldberg & Roswell, 2000). The literature review found that teachers’ experiences of marking and the claims of benefit were varied and depended on context. However, the results of an online survey of Australian teachers broadly echoed the four themes of benefit to teachers Falk and Ort (1998) identified: clarifying their goals and expectations; deepening their discipline knowledge; learning more about students and their work; and developing insights that support their professional practice. The survey results also indicated that there were several minor themes in the Australian teachers’ responses in addition to those four broad categories, including themes of money as an incentive for marking, networking with colleagues and maintaining links to the profession when on parental leave.

Further, the guiding questions for this research invited inquiry into the drawbacks and costs of marking, as well as barriers to entry. The negative aspects of large-scale marking were not found to be much discussed in the literature, but the survey results highlighted stress, anxiety and time pressure as negative impacts. For regional or non-metropolitan teachers, distance was a barrier to taking up marking opportunities, and perhaps this reflects Australian geography. On the whole, though, despite the acknowledged negative aspects of large-scale marking, there were generally positive responses to a question about whether the markers would recommend the experience to colleagues, suggesting the benefits are perceived to outweigh the costs.
The preservice teachers who volunteered for the simulated marking experience mostly had two simple goals, as reported in Chapter 6, and each related to deriving benefit from the simulation experience. The participants sought to increase their confidence and to gain practical experience in marking. When interviewed in groups after the simulation, participants reported feeling more confident and that they valued the experience of scoring authentic student work. Despite seeming different on the surface, the preservice teachers’ perceptions of value in the marking simulation and teachers’ perceptions of value in authentic large-scale marking did intersect. As discussed in Chapter 6, there were broad similarities between the phase one survey responses reported in Chapter 4 and the participants’ experiences reported in Chapter 6. Further, the results from the follow-up phase, reported in Chapter 7, indicated that participants’ positive sentiments towards the simulation and perceptions of having derived benefit from it persisted over the intervening 11 months.

Also, there was evidence of the fourth category of benefit to teachers Falk and Ort (1998) identified: developing insights that support professional practice. There were participants who had adopted some of the processes of the marking simulation, including the training portion, as well as those who had used the rubric in their teaching practice. Further, participants discussed central concepts such as reliability and objectivity in relation to both the simulation and their professional assessment practices. It was pointed out in the discussion section of Chapter 7 that such decisions and reflection on practice may indicate that the simulation supported these preservice teachers in developing their assessment identity (Looney et al., 2018) and/or their professional identity (Feiman-Nemser, 2001). If so, then this may explain how teachers, preservice or not, benefit from marking.

However, there is not enough evidence from this research to argue that teachers and preservice teachers derive the same benefit from large-scale marking – and in fact it is obvious they do not with respect to, for example, monetary remuneration for their time. But there is enough evidence from this research to argue that all the teachers surveyed and interviewed derived some benefit, in diverse and individual ways, from these separate but related practices, and that the general, anecdotal assertion that ‘marking is good PD’ holds true not only for teachers engaging in authentic practice, but also for learners engaging in simulations of practice.

Fidelity, reality and simulations
The researcher’s perspective presented in Chapter 5, though admittedly likely to be biased in unconscious ways, was that the simulation experience was a reasonably faithful replication of
authentic marking practices. Furthermore, based on the analysis in Chapter 4 and the literature review in Chapter 2, there are reasonable grounds to assert that teachers find marking experiences to be beneficial in diverse ways, and Chapter 6 reported that the participants found that simulated marking had benefits for them. Thus, a ‘line’ can be traced between the authentic activity, which is generally considered to be beneficial, to the authenticity of the simulation, through to the perceived benefits to the simulation participants. The metaphorical coherence of this line of reasoning, from authentic practice to an accurate and faithful reproduction of the original, to an equivalently valuable learning experience, is pragmatically appealing because it offers a common sense answer to the question that arose in the course of the researcher’s professional work: will a simulation of the marking practices enacted in my work have similar benefits to the ‘real thing’?

This pragmatic approach, though, does not necessarily consider the experience of the simulation processes. Rather, it is a utilitarian approach that evaluates the simulation based on its outcomes: it appraises the ends of the process, rather than the means. The simplicity of this pragmatic approach is appealing; however, such linear reasoning is reductionist and does not engage with how learning occurs in the simulation experience. This observation led the researcher to consider alternative approaches that would help in understanding the practice more thoroughly, and led to the inclusion of PA theory as an additional lens through which to consider the sayings, doings and relatings within the practice (Mahon et al., 2017; Rönnerman & Kemmis, 2016), which may be the sources of benefit for participants. These sayings, doings and relatings are considered in greater detail in the following paragraphs.

In a simulation, it is generally assumed that one tries to (re)enact the sayings and doings of an authentic experience, but there remains an awareness that the actions are inauthentic, and that one is adopting a persona. This fundamentally alters the socio-political and material-economic arrangements within a simulation, and these interactions make a simulation experience different from the experience that is trying to be recreated, and do so in such a way that participants are always conscious of the façade. For example, authentic marking is done by paid employees at the direction of a manager, which speaks as much to the power dynamics of the social-political arrangements as it does to the material-economic architectures of operational marking. Simulated marking, though, is done by learner-participants and facilitated by an actor drawing on memories and impressions of past experiences. Additionally, the participants in this marking simulation were aware that their doings – their scoring – would have no effect on the students whose work they were judging, thus the simulation markers have no obligation to the producers of the work to be diligent in
scoring. In short, the socio-political and material-economic arrangements, the *relatings* and *doings* of the practice, are fundamentally different.

This is not to say that a simulation cannot replicate large parts of a ‘real’ experience, of course. A marking simulation can emulate aspects of the cultural-discursive arrangements (*sayings*) of authentic marking operations, as well as some aspects of the material-economic (*doings*), and social-political arrangements (*relatings*). The *sayings* include the language used in conducting the training, moderation and marking, and the *doings* include the physical and environmental arrangements including the marking interface and the tools used to mark. The *relatings* are associated with being given the task of reading and judging unknown students’ writing. But even though *sayings*, *doings* and *relatings* can be replicated, a simulation cannot *be* marking: they are two separate practices, even when they share common elements and one is designed to mirror the other.

This problematises the pragmatic line of reasoning described above, where marking has benefits and therefore a faithful replication of marking will provide similar benefits. That reasoning rests on a fulcrum of fidelity, and we can see that even if the outcomes are similar, the processes and experiences to reach those outcomes are necessarily different. However, this has some surprisingly liberating implications for any future iterations of marking simulations.

Hopwood (2017) rebelled against the idea that fidelity is fundamental to simulation, and indeed was bold in proclaiming the need to “shake off an attachment to the notion of ‘fidelity’” (p. 63). Among his key points is that in educational contexts, the pedagogical goals of simulations take supremacy over all other goals, and that simulation offers ways to imagine hypotheticals and explore practice in ways that go beyond reality.

In this research, there was one aspect of the marking simulation that was highlighted in Chapters 5, 6 and 7 as a compromise in fidelity: the conversation among participants during marking. This divergence from an authentic experience is worth considering from the perspective Hopwood offered.

In designing the simulation, the researcher assumed that the main sources of the benefits of marking would be conversations with colleagues and exposure to a range of student work (Gambell & Hunter, 2004; Goldberg & Roswell, 2000; Masters & Forster, 2000). In the researcher’s notes taken during the simulation (see Chapter 5), and also in the post-simulation interviews (see Chapter 6), he noted that he made a conscious decision to not discourage conversation between participants in the marking sessions. Understanding one’s own motivations for a decision is fraught because of the potential for personal bias, but the
researcher justified it to participants in one of the small-group interviews. As reported in Chapter 6, his reasoning was that he refrained from discouraging conversations because the participants were not being paid, did not have explicit targets to meet, and were not expected to be expert markers; essentially, his point was that work conditions are different to simulation conditions, or to use PA terminology, the material-economic and socio-political arrangements differ between the two practices.

Some participants were distracted by the noise level and found it made marking more difficult. Others found that discussion on contentious scores was very helpful. If it is accepted, following Hopwood, that fidelity is not the key to a ‘good’ simulation, then the implication for any future iterations of the marking simulation is that they should be designed according to pedagogical principles, as that is the project at hand. It would have to be decided whether working in silence, as one would in authentic marking, supports the learning objectives of the practice. It was reported in the results section of Chapter 6 that some participants highly valued the moderation exercises, or ‘group marking’ as they called it, over marking in isolation, and that in a similar way, some participants valued the opinion of their peers in scoring the scripts presented to them. So, the researcher suggests that the facilitator of any future marking simulation ought to be responsive to what the participants in a simulation appear to value and need, just as teachers need to in a classroom. For example, in a future iteration it could be made explicit to participants whether a condition, such as silence in the marking room, is to be varied in the simulation. Or, the participants could be consulted on whether they would like the condition varied, or even apportion time to different conditions so that participants can have a range of experiences. As this example shows, once one is freed from the notion of having to faithfully recreate an experience, there are many such possibilities to innovate for the purpose of achieving particular learning outcomes.

**Practical Marking Experience and Initial Teacher Education**

The idea that a marking simulation like the one in this project should become part of ITE courses was first raised and discussed by participants in the post-simulation interviews reported in Chapter 6. Participants raised this idea in relation to the deficiencies they perceived in practicum experiences and assessment units, and the discussion gave rise to a response that has remained with the researcher ever since. It was quoted in Chapter 6, but is reproduced here for the frankness of its expression:

> You can qualify as a teacher and have marked nothing. That’s fucking ridiculous. A lot of people have said that their first prac they marked nothing, a lot of people will
SIMULATED LARGE-SCALE MARKING

mark nothing in their second prac, and we do this theoretical assessment module where we do a lesson plan and we talk about how we’re going to do formative assessment. No-one’s marked anything. (Marker 12)

This indictment, given candidly by a preservice teacher, is noteworthy in many ways; one of those is that this participant feels that a whole cohort of preservice teachers are leaving university underprepared for their work as teachers. There are current tensions about what role universities play in producing “profession-ready” teachers (Charteris & Dargusch, 2018), but that debate is beyond the scope of this research. The point to make here is that participants in this research study reported feeling unprepared for their professional work when they were a semester away from graduating.

This theme of the simulation providing to participants something they did not get in their ITE course persisted into the follow-up phase. In that phase of the research, reported in Chapter 7, the idea of incorporating the experience into ITE courses was raised again, with less vigour but in similarly straightforward ways. Among the examples given in Chapter 7 were Marker 9’s assertions that “I would recommend that it actually gets built into the training, into the teacher training at university, not just for English but especially for English” and “I was really delighted to be offered it, but when I did it, I just thought it should be mandatory.”

As said previously, one of the key goals for the preservice teachers who volunteered was to gain experience – many felt that they simply did not have enough experience in the practical aspects of marking. This is understandable for students who have not yet entered the world of work. One of the key findings of this research, though, is that the participants believed that the marking simulation gave them experience that they did not get through teaching practicums or their course work, and the participants who were interviewed in phase three continued to believe this.

This finding, that participants believed at the time of and well beyond the simulation that it was beneficial enough to become a ‘mandatory’ part of ITE, was unanticipated by the researcher, and yet, he believes, vindicates the research. It provides a clear indication that participants strongly valued the simulation, and that even if the benefits varied from individual to individual, there was general consensus that the benefits were substantial and that there would be similar, significant benefits for other preservice teachers.

Of course, an argument for including such a simulation in an ITE course ought to rely on a more than just the beliefs or perceptions of participants; there ought to be some theoretical justification as well. Feiman-Nemser (2001) developed a continuum of Central
SIMULATED LARGE-SCALE MARKING

Tasks in Learning to Teach which could be used to justify including a simulated marking-experience in an ITE course. There are three stages in the continuum, from preservice, to induction, to continuing PD, each with their own tasks – see Table 16. It is worthwhile considering where the purported benefits of the marking simulation may fit into this continuum.

Table 16 Central tasks of learning to teach. From Feiman-Nemser (2001).

<table>
<thead>
<tr>
<th>Preservice</th>
<th>Induction</th>
<th>Continuing professional development</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examine beliefs critically in relation to vision of good teaching</td>
<td>Learn the context – students, curriculum and school community</td>
<td>Extend and deepen subject matter knowledge for teaching</td>
</tr>
<tr>
<td>Develop subject matter knowledge for teaching</td>
<td>Design responsive instructional program</td>
<td>Extend and refine repertoire in curriculum instruction, and assessment</td>
</tr>
<tr>
<td>Develop an understanding of learners, learning, and issues of diversity</td>
<td>Create a classroom learning community</td>
<td>Strengthen skills and dispositions to study and improve teaching</td>
</tr>
<tr>
<td>Develop a beginning repertoire</td>
<td>Enact a beginning repertoire</td>
<td>Expand responsibilities and develop leadership skills</td>
</tr>
<tr>
<td>Develop the tools and dispositions to study teaching</td>
<td>Develop a professional identity</td>
<td></td>
</tr>
</tbody>
</table>

The participants reported learning about how to make decisions about the quality of student work given and they also reported learning from being exposed to levels of quality of student work they had not seen in their practicums. These experiences are relevant to the central tasks for preservice teachers, as they could be categorised as the participants developing an understanding of learners and learning. Additionally, learning about the practical application of a rubric to make decisions about the quality of student writing is likely to contribute to participants’ beginning repertoires, which includes approaches to assessment (Feiman-Nemser, 2001). Beyond that, participants in this research were in the final year of their preservice training and it may be that there is not a firm division between these stages of learning and that, for example, professional identity is already developing before formally beginning a professional career as a teacher. As Feiman-Nemser observes, “Beginning teachers form a coherent sense of themselves as professionals by combining part of their past, including their own experiences in school and in teacher preparation, with pieces of the present …” (2001, p. 1030). Thus, the confidence and experience reportedly gained by participants, as well as the doings, sayings and relatings that were part of the
practice of the large-scale marking experience, may well have contributed to the participants’ self-image and professional identity as teachers and/or assessors. In this respect, the inclusion of simulated large-scale marking in an ITE course may be argued for not only on the basis of reports from the students and graduates of an ITE course, and also from the theoretical perspective that it offers preservice teachers learning opportunities at an appropriate stage of their education.

**Simulated Large-scale Marking as Professional Development for Preservice Teachers**

So, having synthesised the findings of the three phases of this research into three main findings, we must turn to the question that this research seeks to answer:

To what extent can participation in a large-scale marking simulation provide, for preservice teachers, a professional development experience similar to authentic large-scale marking?

This question depends upon a few presuppositions, from what professional benefits an authentic large-scale marking experience yields, to whether a large-scale marking experience can be simulated. In this chapter so far, the professional benefits have been discussed, as have issues pertaining to simulation and fidelity. These findings address some of the premises the question assumes.

One of the presuppositions of the research question is that the benefits derived from the two experiences could be identified and compared. This research has found that preservice teachers reported two main benefits from participating in simulated large-scale marking: increased confidence and gaining experience. It was found in the survey of teachers and in the literature review that there was a wide range of benefits; confidence and experience certainly seem to be within the scope of reported benefits. However, experienced teachers reported many other benefits such as the usefulness of gaining a bigger picture – a broader frame of reference – in order to understand where their students are at in their learning. Intuitively, such a benefit is outside the possible benefits that student teachers could be expected to derive from the experience, given they do not yet have their own classes to teach. That said, in the follow-up interviews, participants reported that seeing a wide range of quality in student writing had benefits where their practicum experiences had not provided that exposure. As with experienced teachers, then, the benefits the participants received differed among individuals, according to their experiences and contexts, even though two overarching benefits were identified. The conclusion drawn here is that the value reported by
the simulation participants are consistent with the benefits reported by teachers in the literature review and online survey. The nature of the similarity appears to be that simulation participants reported a sub-set of the benefits reported by experienced teachers. So, even though preservice teachers see the benefits of the simulated experience, such experiences can be expected to provide a narrower range of benefits than authentic marking experiences.

Another presupposition of the research question is that a marking experience can be simulated at all. It was found that the fidelity of the simulation in this research was quite strong, based the researcher’s reflective comparison of the simulation to his past experiences of large-scale marking and on quantitative indicators of marking quality. There was also a reflection on the importance of fidelity informed by a perspective discovered in the literature (Hopwood, 2017). That discussion and reflection questioned whether fidelity is necessary for successful pedagogical simulations, and suggested that the lapses in fidelity in the simulation were not likely to have endangered its intended pedagogical outcomes. Taken with the perspectives offered by the researcher, this leads to a conclusion that a large-scale marking operation can be effectively simulated.

So, to what extent did the simulated marking experience provide a PD experience similar to authentic marking? The simulated experience successfully emulated the sayings, doings and relatings of authentic marking. The benefits reported by simulation participants fell into a narrower range, but were consistent with benefits reported by authentic markers. Perhaps most tellingly, both simulation participants and authentic markers recommended the experience to colleagues. On the whole, simulated large-scale marking, much like authentic large-scale marking, has been shown to be ‘good PD’ in the estimation of the preservice teachers who participated.
Chapter 9 – Conclusion

In the literature review in Chapter 2, it was stated that this final chapter would turn to the research framework of Borko (2004). Borko presented a three-phase framework for researching PD, where the first phase is to discover existential proof that “a professional development program can have a positive impact on teacher learning” (Borko, 2004, p. 6). The aims of this research are very well aligned with that aim: it has sought to establish whether a simulation of a large-scale marking operation benefits preservice teachers. Borko (2004) described first phase research into PD as being relatively small, conducted at a single site, and usually designed by the researcher; this research has all of these characteristics.

Additionally, Borko described first phase research as benefiting from using a “multifocal research lens,” which employs “multiple conceptual frameworks and units of analysis” (Borko, 2004, p. 4) to provide a rich account of the processes and participants in the PD activity. These elements are also reflected in the present research. It has been guided by pragmatism, and the researcher selected methods that appeared suitable for answering the questions at hand rather than methods associated with a particular research paradigm or theory. The research has included qualitative analysis of interviews, survey responses and the researcher’s personal records. It has also used some simple quantitative methods to investigate marker agreement and a quantitative exploration of rater severity drawing on Rasch measurement theory. The pragmatic framework itself was supplemented where necessary by brief forays into PA and discussions of identity, which are examples of the situative perspectives that Borko advocates for. All these perspectives taken together make up the ‘multifocal research lens’ of this research.

So, what existential proof has been discovered about the efficacy of simulated large-scale marking? The participants reported a range of benefits, though mostly in gaining experience and confidence, and those perceptions did not wane over the almost-twelve-month gap between simulation and follow-up interviews. Those who participated in the follow-up interviews spoke about using some of the processes, sayings and artefacts from the simulation in their teaching practice. Further, every participant interviewed after the simulation stated that they would recommend the experience to preservice teachers, and several went so far as to advocate that it become part of ITE. There is existential proof in the interview data that the simulation benefited the preservice teachers who participated.

This research also compared the perceptions of the preservice teacher participants with the reports of Australian teachers who had marked in large-scale assessments, and with
the literature about marking as PD. It was found that the benefits the simulation participants discussed were broadly aligned with the benefits reported by teachers who had marked in large-scale operations, though perhaps were narrower in range.

On the whole, this research found that the participants thought the simulated large-scale marking to be beneficial, but such a simple summary is not satisfying. A few of the participants expressed the frustration that preservice teachers felt with their exposure to marking in their teacher preparation to that point (which was near the middle of their final year of their ITE courses); these responses highlighted for the researcher that not only did the participants see the simulation as a good thing to do, but that they thought it so valuable that it ought to be mandatory. The quotations below have been included in previous chapters, but they are included again here as reminders of how strongly some of the participants felt about the experience and its potential role as PD for preservice teachers.

Marker 12 was frank and animated in expressing shock at the lack of marking experience provided in ITE:

You can qualify as a teacher and have marked nothing. That’s fucking ridiculous. A lot of people have said that their first prac they marked nothing, a lot of people will mark nothing in their second prac, and we do this theoretical assessment module where we do a lesson plan and we talk about how we’re going to do formative assessment. No-one’s marked anything. (Marker 12)

While Marker 5 expressed surprise at the absence of training in practical assessment:

I was really surprised, coming into this course, and I don’t know whether you guys feel the same way, but I thought Assessment and Reporting, the unit, was going to be like, how to assess. Not the concepts of assessment for, as and of learning. Do you know what I mean? And, I wasn’t expecting it to be covered in depth but, I thought we would have a class that would be like this. I thought it would be part of our degree. That we’d come in and we’d talk about pulling apart a rubric and working through how to assess people in quite a subject basis. (Marker 5)

It is worthwhile repeating that while these responses are from the post-simulation group interviews, Markers 12 and 5 did not participate in the same group interview – they did not hear each other’s response. The similarity in sentiment, tone and strength of opinion, therefore, is notable.

In the follow-up interviews there were less animated responses, but the sentiment that the simulation was beneficial and should be part of ITE courses was still strong. As Marker 9 said, “I was really delighted to be offered it, but when I did it, I just thought it should be mandatory.” Additionally, one participant, Marker 16, who was interviewed both in a post-simulation group interview and in the follow-up phase, expressed the same idea – that the
SIMULATED LARGE-SCALE MARKING

Simulation ought to be part of ITE courses – in both interviews; this indicated that this perception was enduring.

So, beyond being existential proof that the simulation was effective PD, this research has also provided insight into an area of teacher preparation that these preservice teachers saw as inadequate, at least at this Western Australian university.

Limitations

The primary caveat to this finding is that the simulation was small, with only 22 participants, at one site, with one facilitator. Additionally, the phase one survey also had a small number of respondents. So, whether the results are generalisable is unknown, though nothing in this research has precluded the possibility. Whether different benefits would be derived from running the simulation with a different genre of student writing, or with a different facilitator, or at a different site, or with students at a different point in their ITE, are all open to debate. These caveats are, in fact, avenues for future research, and they articulate into research that Borko (2004) would describe as second phase: taking well-specified PD programs and examining their use in multiple sites, with multiple facilitators and with more participants.

Additionally, as with all qualitative data reporting individuals’ perceptions, there will be limitations. In relying on humans’ recollections and descriptions of their experiences, there will always be unconscious bias, effects of imperfect memory, responses being shaped by social norms and expectations, and a range of other constraints that affect accuracy and objectivity. These limitations are inherent in all survey- and interview-based research, perhaps even all qualitative research.

Significance

Given that this research is but a beginning, raising so many open questions that point towards the need for more research, what can be said of its significance? Primarily, this research has found proof of concept: simulated marking with authentic student work is valued by preservice teachers as beneficial in their ITE. The preservice teachers involved in this research, on the whole, saw it as valuable enough to recommend that it become part of the course they completed. So, this research has significant findings for at least the university that was the site of the research, where these students identified a ‘gap’ in their education and identified an opportunity for the university to enhancing its courses by offering practical marking experiences.
This research also has significance for other universities that are seeking ways to improve their ITE courses; if their courses do not provide preservice teachers sufficient opportunities to gain practical marking experience, then a simulated marking experience is likely to meet this need. It would be likely to be a suitable method for giving preservice teachers additional opportunities to not only mark, but to discuss the qualities of students’ performances and enhance their capacity for evaluative judgment (Carless et al., 2018). It has been already noted that practicum experiences are wildly varied in quality (Rorrison, 2008), and the findings of this research indicate that opportunities for uniform, practical experiences in assessment such as this simulation may well be worth investigating as a way to supplement the workplace learning experiences provided to preservice teachers.

This research has also contributed to knowledge about Australian teachers’ beliefs and perceptions of large-scale marking. Prior to this research, the benefits of large-scale marking were anecdotally known by Australian teachers and used in recruiting material, but there was no research as to whether the perceptions and experiences in Australia are similar to international contexts. While the present research cannot propose a strong argument that this is the case, it has shown consistency between international reports and those of the teachers surveyed in phase one.

Finally, this research has been significant in that it has contributed to the PD of the 22 participants. This is particularly seen in the recollections shared in the follow-up phase by seven of the participants, who had begun teaching in schools. The fact that the marking simulation had given them tools, approaches and strategies that they used in their teaching practice is in itself proof of the benefit of the simulation experience.

**Implications for Future Research**

A final consideration in researching simulation as PD for preservice teachers, or for graduate teachers and teachers who are not able to access opportunities to participate in large-scale marking, is how to move to the second phase of the framework for researching PD developed by Borko (2004). In that phase of research, a program is delivered by multiple facilitators, at multiple sites. This ‘scaling up’ of the PD program would necessitate the consideration of the concept of Fidelity of Implementation (FOI), which is defined by Century, Rudnick and Freeman (2010) as: “The extent to which an enacted program is consistent with the intended program model.” This definition is consistent with earlier work (e.g. O’Donnell, 2008; Ruiz-Primo, 2006).
Century et al. (2010) developed a FOI framework that includes two categories of ‘critical components’ for the implementation of a program. One category is ‘Structural Critical Components’, comprising two critical components: procedural and educative. The other category is ‘Instructional Critical Components’, comprising Pedagogical and Student Engagement critical components. Stains and Vickrey (2017) interpreted this work further and summarised the framework in a diagram that is reproduced as Figure 11. In order for a program to be implemented at multiple sites, these critical components must be specified in detail, as a way of ensuring that the ‘users’ of the program are implementing it in accordance with the intended design.

Figure 11 Categories and subcategories of critical components, from Stains & Vickrey (2017, p. 4).

If this present research were to be implemented at more sites, each of these critical components would need to be specified in detail. That detailed specification is outside the scope of the present research, but some of the findings of this research lend themselves to some speculation about what aspects of this research would need to be included as critical components.

The first critical component that would need to be specified is the use of authentic student writing for participants to mark. This would be a procedural critical component, as the sourcing and curation of a suitable corpus of authentic student work was a fundamental feature of this research. Secondly, an educative critical component would be an experienced facilitator who is familiar with training markers in large-scale assessments, including in explaining rubrics and exemplars, as well as leading moderation discussions. The rubric exemplification and the moderation discussions were all reported by participants to be
beneficial aspects of the simulation. The third critical component suggested by this research would be one of its unanticipated aspects: the discussions between participants as they marked. This would become a student engagement critical component, and participants in future replications of this research ought to be allowed to interact with one another, discussing their judgments of the writing in front of them as they mark.

**Conclusion**

This research arose out of “practical and professional issues and problems” (Punch, 2014). The researcher has worked with many markers, in many contexts, and has heard many times how marking is ‘good PD.’ This research has examined what teachers mean when they say that, and has found that some of those PD benefits can be gleaned from simulating a marking experience. Marking simulation has several practical benefits, including allowing novice raters who may be unreliable to mark without negative consequences for the students whose work they are marking. It also allows for the time and place as well as the fidelity of the experience to be tailored to better meet its learning goals as a PD activity. But well beyond that, this marking simulation has had PD benefits for the participants, providing them with experience in reading, understanding, discussing and judging student work. This gave participants confidence in their ability to be assessors in the classroom, and tools that they used in their first year of teaching.
References


SIMULATED LARGE-SCALE MARKING

Unwin.


SIMULATED LARGE-SCALE MARKING

279). London, UK: SAGE.


184


SIMULATED LARGE-SCALE MARKING

403–419.


Shepard, L. A., Hammerness, K., Darling-Hammond, L., Snowden, J. B., Gordon, E.,
SIMULATED LARGE-SCALE MARKING


Appendix A: Information Letters

Information Letter for Online Survey Participants

The viability of simulated large-scale marking as professional development for preservice teachers

You are invited to take part in a research project that aims to examine whether a simulated large-scale marking session can offer an experience for preservice teachers similar to real large-scale marking. To do so, we first need to ask teachers about their experiences as markers of exams and other large-scale marking.

If you choose to participate, you will be asked to complete a survey. The survey contains questions about your experiences as a marker in large-scale marking operations, including your reflections on the benefits, costs and impacts of being a marker. The questionnaire should take less than 15 minutes to complete.

While there is no direct benefit for you if you agree to complete this survey, you would be contributing in a significant way to what is currently known about the experiences of teachers who participate in large-scale marking, at no risk to you.

Your participation is completely voluntary, and you may stop participating and withdraw at any point, with no consequences, by closing the survey. Your responses will not be recorded until you have completed the survey. You may decline to answer any of the survey questions. All data collected in this survey will be held confidentially and anonymously.

The results of the survey may be used in conference presentations or papers, journal articles and/or a final thesis, but you will not be identifiable in any way in any publications produced from this study. Your data will not be used in any other research projects.

You will have the opportunity at the end of the survey to provide an email address if you would like to receive a summary of the findings at the end of the study.

The ECU Human Research Ethics Committee has approved this study. If you have any concerns or complaints about the research project and wish to talk to an independent person, you may contact:

Research Ethics Officer
Edith Cowan University
270 Joondalup Drive
Joondalup WA 6027
Phone: (08) 6304 2170
Email: research.ethics@ecu.edu.au

If you have any further questions about the research, please contact Nathanael Reinertsen on

Yours sincerely,
Nathanael Reinertsen MEd, GradDipEd, BA W.Aust.
PhD Candidate
School of Education
Edith Cowan University
Initial Letter for Marking Simulation Participants

Dear potential participant,

The viability of simulated large-scale marking as professional development for preservice teachers

I am conducting a PhD research project that aims to evaluate whether a simulated large-scale marking session can provide professional development for preservice teachers, and I am hoping you will consider participating.

What would participation involve?
There are three different parts of the project that you would be volunteering for. Firstly, you will be asked to attend a preliminary interview on campus, in Week 2 of Semester 2. The interview will take no longer than half an hour.

The second, and most substantial, part of the project is a marking simulation. This will involve being trained to apply a professionally developed marking guide to students’ writing, and then marking a number pieces of authentic student work. This portion of the research project is currently planned for Week 3 of Semester 2. The simulation will run from 9.00am to 4.00pm, with lunch provided. At the completion of the workshop you will be given a certificate acknowledging you completion of 6 hours of professional development, addressing the relevant Professional Standards for Teachers in Western Australia (including 5.1 Assess student learning, and 5.3 Make consistent and comparable judgments).

Then, approximately six months after the simulation, you will be contacted to arrange a follow-up interview that will last no longer than half an hour.

Any participation in this research project is completely voluntary, you will be able to withdraw at any point with no consequences.

Giving or withdrawing consent will not affect any relationship with members of the research team or Edith Cowan University, nor will it have any impact on your course assessment.

How do I get involved?
For now, please just note that the first two activities of the research project will happen in the second and third weeks of Semester 2.
In the first week of Semester 2, a more detailed information letter will be circulated that will have a consent form for you to fill out if you wish to participate.

Who do I contact if I wish to discuss the project further?
If you would like to discuss any aspect of this study please contact me, or my principal supervisor Dr Bill Allen.

Yours sincerely,

Nathanael Reinertsen MEd, GradDipEd, BA W.Aust.
PhD Candidate
Edith Cowan University
Tel: [Redacted]
Email: [Redacted]
Information Letter for Marking Simulation Participants

Mr Nathanael Reinertsen  
PhD Candidate  
School of Education  

HREC Approval No: 18155  

Dear participant or potential participant,

The viability of simulated large-scale marking as professional development for preservice teachers

My name is Nathanael Reinertsen and I am conducting a research project as part of my Doctor of Philosophy degree at Edith Cowan University. The research project aims to examine whether a simulated large-scale marking session offers an experience for preservice teachers similar to what teachers report experiencing when participating in real large-scale marking. I am using a prospective longitudinal study methodology where participants in the simulated marking session will be interviewed prior to and after the simulation, and also participate in a follow-up interview approximately six months after the simulation. I am asking final-year education students if they would like to volunteer to participate. This letter sets out information regarding the project and what participation involves.

What does participation in the research project involve?

If you agree to participate in the research, you will be asked to participate in a preliminary interview in at the university campus where you are enrolled. The interview will be no longer than half an hour in length.

The marking simulation is planned for Wednesday the 15th of August 2018, and would require your attendance from 9.00am to 4.00pm. The simulation will include marker training and a four-hour marking session using computers, followed by a small group interview. Lunch and morning tea will be provided.

During the marking simulation, you will read and assess de-identified student work. It is possible that the writing may contain details that could identify the student. By agreeing to participate, you agree to keep all student work that you read confidential.

Approximately six months after the simulation, you will be contacted to arrange a follow-up interview that will last no longer than half an hour.

Are there any potential risks if I choose to participate?

There are minimal risks to you in choosing to participate in this research. During the marking simulation, after training, you will be marking on a computer and be asked to finish your allocation of scripts within the time available. There are some low-level occupational hazards associated with these conditions. Health and safety information, including information about workstation ergonomics as it applies to using computers for marking, will be covered during marker training.
To what extent is participation voluntary, and what are the implications of withdrawing?

Any participation in this research project is completely voluntary, and you may withdraw at any point with no consequences. During interviews you may decline to answer any of the interview questions. You may withdraw from the project after or during any of the interviews or the simulation. Up until the processing and publication of findings, any participant’s data can be withdrawn from the study, so you can even withdraw your consent to participate after participating.

Giving or withdrawing consent will not affect any relationship with members of the research team or Edith Cowan University, nor will it have any impact on your course assessment.

What will happen to the information collected, and is privacy and confidentiality assured?

Any information that could be used to identify you will be permanently removed from the data collected. Any hardcopies of data will be stored securely in locked storage at Edith Cowan University, Mt Lawley. The only people with access to that data will be my two supervisors and myself. Any data collected digitally, or transcribed from hardcopy to digital forms, will be stored in encrypted files, on password-protected computers with up-to-date antivirus software installed. The data will be stored for 5 years and then will be permanently destroyed by shredding hard copies of interviews and notes and deleting computer files.

The identity of participants will not be disclosed at any time, with the sole exception being where there is a legal requirement to do so. Participant privacy, and the confidentiality of information disclosed by participants, is assured in all other circumstances.

Data collected in this study will only be used this project; it will not be used in any extended or future research without first obtaining explicit written consent from participants.

Is this research approved?

The research has been reviewed and approved by the ECU Human Research Ethics Committee.

Who do I contact if I wish to discuss the project further?

If you would like to discuss any aspect of this study please contact me, or one of my supervisors:

Dr. Bill Allen
+61 8 6304 6729
w.allen@ecu.edu.au

A/Prof Brian Moon
+61 8 6304 6275
b.moon@ecu.edu.au

If you have any concerns or complaints about the research project and wish to talk to an independent person, you may contact:

Research Ethics Officer
Edith Cowan University
270 Joondalup Drive
JOONDALUP WA 6027
(08) 6304 2170
research.ethics@ecu.edu.au
How do I indicate my willingness to be involved?
If you have had all questions about the project answered to your satisfaction, and are willing to participate, please complete the consent form on the following page and return it to Nathanael, Bill or Brian.

This information letter is for you to keep.
Yours sincerely,
Mr Nathanael Reinertsen MEd, GradDipEd, BA W.Aust.
PhD Candidate, ECU School of Education
Tel: [REDACTED]
Email: [REDACTED]
Information Letter for Follow-up Interviews

Mr Nathanael Reinertsen  
PhD Candidate  
School of Education

HREC Approval No: 18155

Dear participant,

I am writing to you because you participated in a marking simulation that was held in August last year and you provided your consent and an email address so that I could contact you to arrange a follow-up interview.

I am hoping to interview participants who are now working as teachers, whether that is full-time, part-time or casual relief. The purpose of the interview is to collect impressions and reflections on the marking simulation experience and to ask if you have found any parts of the experience relevant or useful to your work as a teacher.

The interview would last approximately half an hour, and can be conducted at a place and time of your choosing (including on campus at ECU or via teleconference) between the 8th and 19th of July.

As with your past participation in the research project, participation is completely voluntary; you can decline to participate, and if you agree to participate you will be able to withdraw from the research at any point with no consequences.

If you would like to ask any questions or discuss any aspect of this research please contact me, or my principal supervisor, Dr. Bill Allen, +61 8 6304 6729, w.allen@ecu.edu.au

This research has been approved by the ECU Human Research Ethics Committee.

If you have any concerns or complaints about the research project and wish to talk to an independent person, you may contact:

Research Ethics Officer  
Edith Cowan University  
270 Joondalup Drive  
JOONDALUP WA 6027  
(08) 6304 2170  
research.ethics@ecu.edu.au

If you consent to participating in a follow-up interview, please complete the consent form attached and return it to me via email.

Yours sincerely,
Nathanael Reinertsen  MEd, GradDipEd, BA W.Aust.
PhD Candidate  
Edith Cowan University  
Tel:  
Email:  

195
Appendix B: Consent Forms

Consent Form for Online Survey

The viability of simulated large-scale marking as professional development for preservice teachers

Thank you for your interest in participation in this research project, and taking the time to read this information.

This research project aims to examine whether a simulated large-scale marking session can offer an experience for preservice teachers similar to real large-scale marking. To do so, we first need to ask teachers about their experiences as markers of exams and other large-scale marking.

If you choose to participate, you will be asked to complete a survey. The survey contains questions about your experiences as a marker in large-scale marking operations, including your reflections on the benefits, costs and impacts of being a marker. The questionnaire should take approximately 15 minutes to complete.

Your participation is completely voluntary, and you may stop participating and withdraw at any point, with no consequences, by closing the survey page. Your responses will not be recorded until you have completed the survey. You may decline to answer any of the survey questions. All data collected in this survey will be held confidentially and anonymously.

The results of the survey may be used in conference presentations or papers, journal articles and/or a final thesis, but you will not be identifiable in any way in any publications produced from this study. Your data will not be used in any other research projects.

You will have the opportunity at the end of the survey to provide an email address if you would like to receive a summary of the findings at the end of the study.

I agree that by continuing, and submitting the online survey using the arrows below, I give my consent for the results to be anonymously recorded and used in this research.
# Consent Form for Marking Simulation

**Consent Form**

- I have read and understood the information letter about the project.
- I have taken up the invitation to ask any questions I may have had, and am satisfied with the answers I received.
- I understand that participation in the project is entirely voluntarily.
- I am willing to become involved in the project, as described.
- I give permission for interviews that I participate in to be recorded using audio and/or audio-visual recording equipment.
- I understand I am free to withdraw that participation at any time.
- I understand I can withdraw my data up from this study at any time prior to December 2018.
- I understand that the student work that I mark during the marking simulation must be kept confidential.
- I give permission for my contribution to this research to be presented at conferences, published in journal articles and online, provided that I am not identified in any way.
- I understand that I can request a summary of findings once the research has been completed.

| Name of Participant (printed): | _________________________________ |
| Signature of Participant:      | _________________________________ Date: / / |
| Student email address:         | @our.ecu.edu.au |

*Please return the completed form to Dr. Bill Allen, School of Education.*
Consent Form for Follow-up Interviews

Consent Form

- I have read and understood the information letter about the follow-up interview for this research project.

- I have taken up the invitation to ask any questions I may have had, and am satisfied with the answers I received.

- I understand that participation in the project and this interview is entirely voluntary.

- I am willing to be interviewed, as described in the information letter.

- I give permission for the interview that I participate in to be recorded using audio and/or audio-visual recording equipment.

- I understand I am free to withdraw that participation at any time.

- I understand I can withdraw the data I provide in my interview from this study at any time prior to December 2019.

- I give permission for my contribution to this research to be presented at conferences, published in journal articles and online, provided that I am not identified in any way.

- I understand that I can request a summary of findings once the research has been completed.

Name of Participant (printed):

Signature of Participant: ___________________________ Date: / /

Preferred date, time and location for interview:

Date:

Time:

Location: ___________________________
Appendix C: Phase One Questionnaire Items

1) Are you a current or former schoolteacher in Western Australia?  
   Current / Former / Other: [Open Response]

2) How many years’ teaching experience do you have in total?  
   [Numerical Response]

3) How many years’ experience do you have teaching in Western Australia?  
   [Numerical Response]

4) Have you ever participated in large-scale marking (for example TEE or WACE exam marking, or NAPLAN marking)?  
   Yes / No

   If ‘Yes’ to Q4:  
   5) How many times have you participated in large-scale marking? [Numerical Response]

   6) Do you intend to participate in any large-scale marking in the next 12 months? Yes / No

   ➔ Go to Question 10

   If ‘No’ to Q4:  
   7) Have you ever considered participating in large-scale marking?  
      Yes / No

   If ‘Yes’ to Q7:  
   8) What has prevented you from participating in large-scale marking?  
      [Open Response]

   If ‘No’ to Q7:  
   9) Why have you never considered participating in large-scale marking?  
      [Open Response]

   ➔ Go to End of questionnaire
10) Listed below are some statements about how participating in large-scale marking may affect teachers.

Select ALL the statements below that you agree with.

Marking a large-scale assessment:
clarified the learning goals and expectations you have of your students. ☐
[distractor] increased your repertoire of teaching strategies ☐
deepened your discipline knowledge. ☐
taught you more about students and their work. ☐
supported your professional practice. ☐
[distractor] developed your sense of professionalism ☐
raised your awareness of students’ literacy behaviours. ☐
increased your confidence in your ability to judge the quality of student work. ☐
deepened your knowledge of the assessment itself. ☐

[Instruction] ‘The next two questions are about what you can remember about the first time you marked a large-scale test.’

11) Why did you decide to apply to be a marker?
[Open Response] / Can’t recall

12) Did you have any expectations about the marking experience?
[Open Response] / Can’t recall

13) Finally, if another teacher asked you whether he or she should apply to be a marker, what advice would you give?
[Open Response]
## Appendix D: Phase One Codebook

### Nodes

<table>
<thead>
<tr>
<th>Name</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advice for potential markers</td>
<td>85</td>
</tr>
<tr>
<td>Affirmative</td>
<td>21</td>
</tr>
<tr>
<td>Emphatically Positive</td>
<td>14</td>
</tr>
<tr>
<td>Gives insight into students</td>
<td>2</td>
</tr>
<tr>
<td>Minimum necessary experience</td>
<td>1</td>
</tr>
<tr>
<td>Networking</td>
<td>2</td>
</tr>
<tr>
<td>Neutral</td>
<td>8</td>
</tr>
<tr>
<td>Pay not worth it</td>
<td>1</td>
</tr>
<tr>
<td>PD Benefits</td>
<td>25</td>
</tr>
<tr>
<td>Clarification</td>
<td>6</td>
</tr>
<tr>
<td>Deepen Understanding of Students</td>
<td>1</td>
</tr>
<tr>
<td>Improve Accuracy</td>
<td>1</td>
</tr>
<tr>
<td>Improve own assessments</td>
<td>2</td>
</tr>
<tr>
<td>Increase Confidence</td>
<td>1</td>
</tr>
<tr>
<td>Increase marking speed</td>
<td>1</td>
</tr>
<tr>
<td>Vague</td>
<td>9</td>
</tr>
<tr>
<td>Warning about time</td>
<td>10</td>
</tr>
<tr>
<td>Wider perspectives</td>
<td>1</td>
</tr>
<tr>
<td>Barriers to marking</td>
<td>25</td>
</tr>
<tr>
<td>Distance</td>
<td>2</td>
</tr>
<tr>
<td>Lack of confidence</td>
<td>2</td>
</tr>
<tr>
<td>Lack of experience</td>
<td>2</td>
</tr>
<tr>
<td>Lack of opportunity</td>
<td>2</td>
</tr>
<tr>
<td>Other barriers</td>
<td>6</td>
</tr>
<tr>
<td>Heard negative things</td>
<td>1</td>
</tr>
<tr>
<td>Other ways to get similar PD benefits</td>
<td>1</td>
</tr>
<tr>
<td>Tiresome or Monotonous</td>
<td>1</td>
</tr>
<tr>
<td>Washback of large-scale assessment</td>
<td>1</td>
</tr>
<tr>
<td>Time</td>
<td>11</td>
</tr>
<tr>
<td>Expectations of marking</td>
<td>57</td>
</tr>
<tr>
<td>Beneficial</td>
<td>8</td>
</tr>
<tr>
<td>Challenging</td>
<td>3</td>
</tr>
<tr>
<td>Difficult</td>
<td>7</td>
</tr>
<tr>
<td>Insider's Perspective</td>
<td>1</td>
</tr>
<tr>
<td>Interesting</td>
<td>1</td>
</tr>
<tr>
<td>Money</td>
<td>1</td>
</tr>
<tr>
<td>Negative emotional reactions</td>
<td>8</td>
</tr>
<tr>
<td>Intimidating</td>
<td>2</td>
</tr>
<tr>
<td>Nervous</td>
<td>4</td>
</tr>
<tr>
<td>Pressure</td>
<td>2</td>
</tr>
<tr>
<td>Networking Opportunity</td>
<td>2</td>
</tr>
<tr>
<td>No expectations</td>
<td>9</td>
</tr>
</tbody>
</table>
## SIMULATED LARGE-SCALE MARKING

<table>
<thead>
<tr>
<th>Reason for marking</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process specific expectations</td>
<td>3</td>
</tr>
<tr>
<td>Time Consuming</td>
<td>11</td>
</tr>
<tr>
<td>Tiresome or monotonous</td>
<td>3</td>
</tr>
<tr>
<td>Gain a Wider Perspective</td>
<td>8</td>
</tr>
<tr>
<td>Gain an insider's perspective</td>
<td>11</td>
</tr>
<tr>
<td>Gain experience</td>
<td>3</td>
</tr>
<tr>
<td>Interest</td>
<td>1</td>
</tr>
<tr>
<td>Invited</td>
<td>1</td>
</tr>
<tr>
<td>Job expectation</td>
<td>1</td>
</tr>
<tr>
<td>Maintain links to the profession while being a primary carer</td>
<td>3</td>
</tr>
<tr>
<td>Money</td>
<td>11</td>
</tr>
<tr>
<td>Networking</td>
<td>3</td>
</tr>
<tr>
<td>Professional Development</td>
<td>23</td>
</tr>
<tr>
<td>Become a better teacher</td>
<td>7</td>
</tr>
<tr>
<td>Deepen understanding of their subject</td>
<td>5</td>
</tr>
<tr>
<td>Deepen understanding students</td>
<td>1</td>
</tr>
<tr>
<td>Increase confidence</td>
<td>1</td>
</tr>
<tr>
<td>Increase Skill or Competence</td>
<td>4</td>
</tr>
<tr>
<td>Recommended by colleague</td>
<td>1</td>
</tr>
</tbody>
</table>
Appendix E: Simulation Marking Interface

The home screen presented to markers when they logged in:

![Home Screen](image)

The marking interface, showing a placeholder student response with the marking criteria labels and radio buttons for assigning scores:

![Marking Interface](image)
Dear <First Name>,

Once again, thank you for participating in my PhD research. As promised, I have done some analysis of the marking data and this email is to give you some idea of how your scoring compares to the scores that were given by professional markers to the same scripts you marked. Please feel free to contact me if you would like more information or have any questions.

You logged in as <Marker_##>, and marks were recorded for a total of <x> scripts for this log-in. The average number of scripts marked by other participants in the simulation (not including those who had technological issues with their log-in) was 26.

The correlation between the total scores of scripts you marked and the total scores from trained markers was <y>. As a general rule of thumb, anything over 0.7 is considered a reasonably strong correlation – so well done!

I also calculated the percentage agreement rates for each criterion score you gave (there were nine criteria scores for each script you marked). <z>% of the time, your mark was exactly the same as the score from the professional marker, and another <q>% of the time your mark was one mark away from the score from the professional marker. Ideally, marker supervisors look for exact plus adjacent agreement to be approaching 100%.

Congratulations on a job well done!
Appendix G: ConQuest Command Files for Rasch Analysis

First Run with panel parameter

```plaintext
Title AllData_ByPanel;
datafile AllData.dat;

format panel 1 rater 2-3 pid 4-7 responses 8-16 ! criteria (9);
labels << eWriteCrit.lab;
codes 4,3,2,1,0;
model rater + criteria + criteria*step + panel;
export xsi >> myParams_1.txt;
set constraint = cases;
estimate ! nodes = 50;
show ! estimates=latent >> Out_1.shw;
itanal >> Out_1.itn;
```

Second Run without panel parameter

```plaintext
Title AllData Run 2;
datafile AllData.dat;

format panel 1 rater 2-3 pid 4-7 responses 8-16 ! criteria (9);
labels << eWriteCrit.lab;
codes 4,3,2,1,0;
model rater + criteria + criteria*step;
export xsi >> myParams_2.txt;
set constraint = cases;
estimate ! nodes = 50;
show ! estimates=latent >> Out_2.shw;
```
Appendix H: Phase Two Interview Schedules

On the following two pages are the questions that were used to guide the pre- and post-simulation interviews in Phase Two. The questions are not a complete list of exact questions asked, rather they were a guide for the structure of the interviews with the aim of ensuring each interview covered similar ground, while allowing the researcher to ask additional probing questions.

The interview guiding questions, those in the middle columns, were all designed to elicit responses that will be relevant to the fourth guiding research question: “4. What do participants in a simulation of large-scale marking report about the experience?” The possible probing questions in the last column were prompts for probing questions.
# Pre-simulation Interview Schedule

<table>
<thead>
<tr>
<th>Research Guiding Question</th>
<th>Guiding Questions</th>
<th>Probing questions</th>
</tr>
</thead>
</table>
| 1) What motivated you to volunteer for this experience? | • What was your main reason for volunteering?  
• Did you have to weigh participation against other commitments or activities? If yes, why prioritise this activity over others?  
• Did you discuss participating with others? To what extent did their response affect your choice? | |
| 2) What expectations do you have of the experience? | • What, if anything, are you hoping to “get out of” today?  
• Do you expect marking to be hard? In what way?  
• Do you have any reservations about participating? | |
| 3) Can you identify what has informed or help to create your expectations? | • Did you have discussions with peers/mentors/family?  
• Have you seen marking referred to as a professional development activity? If yes, what do you recall about the context in which it was referred to and what was mentioned about it? | |
| 4) What outcomes do you hope to achieve from the experience? | • Do you think there will be positive outcomes? What?  
• Do you think there may be negative outcomes? What? | |
<table>
<thead>
<tr>
<th>Research Guiding Question</th>
<th>Guiding Questions</th>
<th>Possible probing questions</th>
</tr>
</thead>
</table>
| What do participants in a simulation of large-scale marking report about the experience?” | 1) Are you feeling generally positive or generally negative about the marking experience you’ve just had? | What do you feel positive about?  
What do you feel negative about? |
|                           | 2) How did the experience match with your expectations?                           | Was there anything you were hoping to “get out of” today that you don’t think you did?  
Were there things you were not expecting? If yes, what was your reaction to it/them? |
|                           | 3) What parts of the marking session were the most difficult?                      | Did the difficult aspects become easier as you the day progressed?  
Did anything become more difficult as the day progressed? |
|                           | 4) What parts of the marking session were the easiest?                            | Were they easy because you had done similar things before?                                |
|                           | 5) Would you recommend participation in similar experiences to peers or colleagues in future? | Why/Why not?                                                                                |
Appendix I: Phase Three Interview Schedule

Below are the questions that were used to guide the interview of participants in Phase Three of the research. The questions were not intended to be a complete list of exact questions to be asked, but rather a guide for the structure of the interviews to ensure each individual’s interview covered similar ground, while allowing the researcher to ask additional probing questions.

The interview guiding questions, those in the middle columns, were all designed to elicit responses that will be relevant to the seventh guiding research question: “7. After having transitioned from preservice to early career, will the participants’ impressions of the experience have changed compared to those captured immediately after the marking session?” The possible probing questions in the last column were used as prompts.

<table>
<thead>
<tr>
<th>Research Guiding Question</th>
<th>Guiding Questions</th>
<th>Possible probing questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>After having transitioned from preservice to early career, will the participants’ impressions of the experience have changed compared to those captured immediately after the marking session?</td>
<td>1) How well do you feel you remember the marking simulation?</td>
<td>• Do you think of it when you assess student work?</td>
</tr>
<tr>
<td></td>
<td>2) Do you ever think of that experience in the context of your teaching?</td>
<td>• Does it come up in conversations with your colleagues?</td>
</tr>
<tr>
<td></td>
<td>3) Do you think the simulation has had any lasting impacts?</td>
<td>• Positives? • Negatives?</td>
</tr>
<tr>
<td></td>
<td>4) Would you recommend participation in simulated marking experiences to peers or colleagues in future?</td>
<td>• Were they easy because you had done similar things before?</td>
</tr>
<tr>
<td></td>
<td>5) Do you feel that the simulated marking session has made you more or less likely to pursue marking opportunities?</td>
<td>• Why/Why not?</td>
</tr>
</tbody>
</table>