# How can I be sure?: Revisiting Assessment Practices in GCSE English in the FAVE Sector

Michael Smith

**For PhD**

A thesis submitted in partial fulfilment of the requirements of the University of Sunderland for the degree of Doctor of Philosophy PhD

May 2020

# Contents

# Abstract

This thesis addresses the question of how assessment practices can be better understood in relation to individual teacher interpretations of subjective criteria. To achieve this, this research study considers the possible benefits and challenges of using an adaptive comparative judgement (ACJ) approach to the summative assessment of GCSE English students' creative writing in the FAVE (Further, Adult and Vocational Education) sector.

The extent to which an ACJ approach to assessment can strengthen the fairness, accuracy and integrity of assessment judgements are explored as well as the value of collaborative working and the sharing of assessment judgements in relation to standards, content and quality of students' work. The research adopts a mixed methods approach, including interviews, to provide insight into teachers' use of an adaptive comparative judgement approach to assessing creative writing text quality. A justification for viewing these findings through an interpretivist paradigm is advocated, which it seen as central to understanding the nature of this assessment practice.

Much of the discussion centres on ideas of what is meant by good quality, professional expertise, relative educational value within assessment practices, and what makes a judgement correct and meaningful. Findings suggest that teachers draw on internalised quality markers that exist in tacit form when assessing through adaptive comparative judgement, and that a collaborative and dialogic approach to the understanding and sharing of these is crucial if high quality assessment practice is to be fostered and maintained.

**Keywords:** GCSE English, creative writing, assessment practice, tacit knowledge, adaptive comparative judgement

# Chapter 1: Introduction

## Context and Problem

### Context: Definitions and perceptions of assessment in General Certificate in Secondary Education (GCSE) English

Although much has been written about assessment theory and practice from a range of different perspectives, until recently, it has been much less easy to find literature which challenges the purposes of assessment, the qualities it does/should identify and matters of teacher judgment. Innovations in assessment theory and practice often reflect the dominant political ideology of their time, so if we want to know the real purpose of an education system it is wise to look into its assessment procedures as these can offer insights into the inexorable links between any given assessment theory and the political, social and institutional contexts in which it is applied. There exists little or no common ground in which debates and discourses surrounding 'effective' assessment can be anchored, such is the dependence on context in informing and defining assessment practices. Although definitions differ, we can look to Hoy and Hoy's (2013) assertion that assessment is a "process of gathering information about students' learning." (2013: 263). This Chapter discusses two prominent methods of assessment that teachers make use of on a regular basis: formative and summative assessment. These discussions feature considerations of what contemporary educational discourses tells us about these modes of assessment, and their relative educational value in wide-ranging, complex and unfolding situations. This Chapter goes on to elaborate on the enduring educational issue that this thesis seeks to address, namely, the notion that assessment practice

involves making a judgement that requires nuanced and tacit understanding and implicit knowledge of a myriad of contextual features. It argues that without this, the central tenets and guiding principles of both formative and summative modes of assessment can be neglected by teachers.

Formative assessment has for many years been recognised as a critical facet of the effective teacher's repertoire. It can be conceptualised as a pedagogy of contingency, in which information interpreted and elicited by the teacher from the student is used to influence his or her learning through tailored instruction that is contingent on what the student already knows (Natriello, 1987; Crooks, 1988; Kluger and DeNisi, 1996; Black & Wiliam, 1998; Nyquist, 2003; Wiliam, 2006). As formative assessment can be conducted by teachers with little preparation required, it often operates as a non-invasive, informal and low stakes form of assessment. Questions posed by teachers in class are a useful example of a commonly used formative assessment strategy: effective teachers might pose as many as one question every two minutes, and between fifty and one-hundred questions per hour in class (Hastings, 2003), some directed to individuals and some to the group, each question posed with the intention to provide the teachers with insights into what their students are thinking. Formative assessment strategies can be deployed by the teacher in a discretionary fashion informed by the detailed knowledge of each student being concerned, so that students can be afforded as many opportunities as possible to demonstrate their knowledge in any given context.

However, across the educational landscape formative assessment is not the only kind of assessment worthy of systematic research and development. Summative

assessment serves a different but no less valuable purpose when compared with its formative counterpart. Wiliam defines summative assessment as something that takes place after learning has occurred in order to quantify student performance against a specific measure: ‘if you’re assessing in order to grade students, to rank them or to give them a score on a test, then that’s assessment of learning” (2006:7). The idea of ranking and classifying student performance through measurement is not a new one, and is one that certainly predates formative assessment, as outlined above, by several decades at least. However, there has in recent years been a significant sea change in the perception of summative assessment. In some circles, the idea of measuring learning via an imposed judgement against a set of standards has led to summative assessment becoming vilified and demonised particularly in the field of education to the extent that formative assessment now dominates much of the educational discourse, with summative assessment relegated to perceived position being of lesser educational value.

In effect, summative assessment, through the application of a kind of perverse ‘technical- rational’ logic (Dunne, 1993), has been reduced to the status of the only most instrumental and narrow of assessment practices solely for the purposes of comparison and accountability. Teachers, colleges, boroughs, and nations are required to publicly report and share the performance of their students in specific assessment tasks with frequency and consistency. Such demands can inevitably change the lens of the focus of assessment for teachers and institutions, and can have direct and unintended consequences for the quality and breadth of what is widely taken to be ‘good’ education. Coffield (2008) encourages teachers to question their own stance towards the creeping tendency of teaching to the test, by asking:

‘Do we require our learners to think for themselves or just to report other people’s thinking? Do we teach them how to find and pose problems as well as solve them? Are they regurgitating ‘unwanted answers to unasked questions’ just to pass exams?’ (2008:29). One might argue that concerns such as these are well-founded; research from the field of sociology by Dorling (2015) highlights that the United Kingdom’s reliance on test-centric teaching leads to an inability in fostering long term understanding and deeper levels of learning in school leaving students when compared with international counterparts (2015:6). Dorling suggests that these ideas have permeated the education landscape on a national scale. In view of this, we can understand how summative assessment has come to be considered synonymous with ideas of performativity, reductive teaching and curriculum models that privilege educational outcomes over educational processes.

In considering the above, we can draw quite stark contrasts between the two modes of assessment; formative as a dynamic and flexible approach that can be deployed by teachers to help inform them of their student’s future learning trajectory, and summative as an evaluative and rigid judgement that seeks to capture learning that has already taken place. Alternative but commonly used expressions for these concepts seem to cement this polarity further. Formative assessment is recognised as Assessment *for* Learning (AfL), and summative assessment as Assessment *of* Learning (AoL). These terms seem to imply that formative assessment helps students to learn *in the event*, and summative assessment helps to determine if learning has taken place *after the event*. To sustain this somewhat polarised line of thinking, it is assumed that there is a degree of finality to AoL, as if once the end of a learning episode has been reached there is no need to consider *how* the student

might improve as they are no longer *in the event* of learning; the window of opportunity for a pedagogy of contingency to be employed by the teacher based on what the student already knows has since passed. But the reality is that formative and summative assessment are not two static concepts that exist poles apart on opposing sides of a spectrum, and reaffirming this notion by subscribing to the idea that summative assessment cannot lead to valuable learning experiences in a manner not dissimilar to formative assessment is not only short-sighted but also rather dangerous.

Let us consider the following three scenarios:

1. A student studying on a Beauty Level 1 programme is conducting a consultation with a client in the training salon. He converses with his client and fills out a consultation form with due care and attention. As he does this her tutor observes him. After the consultation has taken place, the tutor completes an observation record sheet and provides the student with written and verbal feedback on her performance in the activity.
2. A student studying on a BTEC Level 3 Subsidiary Diploma in Sport completes an end of unit assignment for formal submission to her teacher. She submits her work and receives detailed feedback four days later on her performance against the unit learning objectives, some of which map to other criteria she has yet to be formally assessed against.
3. A student studying GCSE English Language completes a mock exam on week twelve of her thirty-two-week course. She completes it in a large hall alongside her peers under exam conditions. The paper is assessed by her

GCSE English teacher; each question is scored and her paper is given a total and graded, and she is then given detailed written feedback suggesting where she might improve in the future.

At first glance each of these scenarios appears to include an example of summative assessment taking place. We see student learning being judged and rated against what we can assume is a pre-set set of criteria and standards by their teacher. Furthermore, these assessments seem to be seeking to identify if the student has retained and can demonstrate learning that they have acquired as a result of prior experiences - an assessment *of* learning. But strikingly each of these examples also includes the student receiving detailed feedback from their teacher. There is a very real chance that these three students will have deepened their understanding of their subject as an outcome of the assessment and resulting feedback, perhaps even more so than if it were administered as a formative activity. Crowley (2010) observes that summative assessments often coincide with a sudden boost in student motivation, and that students value the opportunity to have their understanding of a topic formally assessed through such processes.

The FAVE (Further, Adult and Vocational Education) sector comprises many subjects and qualifications that feature modularised rather than linear delivery and permit the flexible positioning of summative assessments throughout an academic year rather than scheduling all assessments towards the end of a course. In such instances the teacher plays a critical role, in that they are also acting as the student's assessor. Unlike qualifications in which summative assessment takes place by a neutral third party external to the institution, as with many end-of-year and online

examinations, each of the scenarios presented above feature teachers making summative judgements as to their own students' learning. As a result, teachers such as those featured in the scenarios above that are summatively assessing units of assessment prior to the end of the programme are in a position to adopt pedagogies of contingency as they continue to teach these students over the course of the academic year. Of course, such contingencies might run the risk of being born solely of the student's performance in the assessment and may not consider a more holistic appreciation of the student's learning. This is perhaps where we can locate Coffield's concerns regarding 'unwanted answers to unanswered questions' (2008:29). However, a much more favourable outcome will be for the teacher to be equipped with the requisite skills, knowledge and experience to reach judgements that can offer both an assessment of learning, and an assessment for learning on an iterative basis. What we can conclude at this juncture is that assessment of both varieties fluctuates in educational value for both student and teacher depending on the way in which it is applied in a given context.

Through examination of the scenarios above, this thesis identifies possible ways in which summative assessment can serve to positively impact on student learning. The characterisation that summative assessment is only concerned with identifying what learning has taken place after the act is reductive and ignores learning that can occur as a result of the assessment itself, and any subsequent feedback that might be forthcoming. The misrepresentation of summative assessment as being detached from the learning process as a result of its quantification of learning is a highly questionable one, and is something of which teachers must be cognisant. With this broader definition of summative assessment now explicated, we can begin to explore

possibilities in relation to how this practice might be improved and refined to best operate in the contexts in which it is located. The following discussion elaborates on the specific subject context in which this research enquiry is based. It highlights key factors and influences that hold significant sway in how, when and why the assessment practices aligned to the teaching of GCSE English are conducted in the manner in which they are in a large general Further Education College in England which forms the site of this study.

**GCSE English as a subject in the FAVE sector**

English as a subject has endured a turbulent history within the Further Education sector, particularly in recent times. In the last two decades alone, we have seen political reforms, and subsequent funding allocations, that have shifted the sector from delivering Basic Skills (2001) to students predominantly in a one-to-one, individualised mode, to Key Skills (2004), and then to their more worldly cousin Functional Skills (2010), in which students learned skills that map to the real-life application of these subjects, often in group settings. These changes alone chronicle the significant upheaval and policy storms that Further Education students, teachers and institutions have weathered. The Wolf Report (2011) heralds the most significant reform to date, with the recommendation that 'Students who are under 19 and do not have GCSE A*-C, or grade 4 in English and/or Maths should be required, as part of their programme, to pursue a course which either leads directly to these qualifications, or which provides significant progress towards future GCSE entry and success' (Wolf, 2011:15). These recommendations were committed to policy with the Government of the day's *Maths and English provision in post-16 education* (2014), a

written statement, adopted by colleges in the UK at the commencement of the 2015-16 academic year. It is this adoption of GCSE English Language as a target qualification for 16-18 students studying the Further Education sector, and the resulting assessment practices that accompany it, that will form much of the focus and discussion in this study.

Justifications for the adoption of GCSE English in Further Education settings are worthy of exploration at this point. Wolf (2011) highlights that a DFE review which examined a cohort of young people who were 15 in 2005/6 and studying on vocational courses, established that 'the percentage of the cohort with both maths and English GCSE A*-C rises from 44.8% at 15 to 49% at 18 – still below half, and less than a five percentage point rise' (Wolf, 2011: 83). This stagnation was blamed on Key Skills, which were often delivered without specialist instruction, did not feature writing at all in on-demand tests, instead offering students multiple choice as opposed to open-ended questions, and consequently provided no sense of equivalence with GCSE grades to both students and employers (Wolf, 2011; Fuller & Unwin, 2011). Explicating the reasons for the adoption of GCSE qualifications in Further Education are worthwhile, as they offer an insight into where previous qualifications have faltered and failed to provide the educational outcome desired by the government of the time. Clearly, and unsurprisingly, student success and progression rates are at the top of the agenda here. Moreover, there is an acknowledgement that subject specialists are required to teach English rather than vocational teachers. Finally, we can observe that a broader curriculum that also assesses student writing is necessary. It is with an appreciation of these motivations, namely: the desire for higher success rates, the need for a more knowledgeable and

capable teaching workforce and a broader curriculum, that we can locate our discussion of the practicalities of assessing GCSE English from the perspective of a teacher based in the FAVE sector.

**The context for this research**

This research is based at a General Further Education College based in North-East London. The college is a provider of vocational qualifications across a range of vocational and academic subjects. This include those situated in construction and trade, digital, creative, health and science and service industries. All students at the college between the ages of 16-18 continue to study English in some capacity alongside their chosen vocational qualification, as per recommendations from the Wolf Report (2011). The college has a cohort of approximately 3,000 16-18 students. In addition to this 16-18 student cohort, the college also has a considerable adult enrolment (19+) across a range of full-time and part-time programmes.

The college has ten full-time teaching staff who contribute to the delivery GCSE English across varying modes of study at the college. In what can be considered to be an indicative trend for the wider Further Education sector, only one of these ten teachers had some experience of teaching GCSE English Language before the shift in policy in 2015 extended the qualification's scope to include full-time students without an A*-C grade currently studying in the FAVE sector. The remaining nine teachers had originally been employed as teachers of Functional Skills English, which as has been established, is an entirely different curriculum serving a very different purpose. The college saw an increase from 112 GCSE English students in the 2014-15 academic year to 668 and 712 GCSE English students in the 2015-16

and 2016-17 academic years respectively. Again, this trend is not exclusive to this college. Other UK Further Education institutions have faced similar increases in student cohort size in light of the aforementioned reforms. Across the UK 59,558 students aged between 16-18 we resubmitted for examination for a GCSE English Language qualification in the 2016/17 academic year as part of their continuing studies in Further Education (FEWeek:2017).

In view of such changes, there remains an operational as well as moral obligation to impart the very best education to our students, and to provide learning opportunities of the highest quality to them. Circumstances do little to assist with this, however. Colleges have one academic year in which to teach the GCSE English Language curriculum to students, and not the two years that schools have available. With only one year available to teach the curriculum, the role of initial and diagnostic assessment on English programmes is of paramount importance to help inform a student's 'learning trajectory' (Crowley, 2010, Roberts & Smith, 2014), However, such practices are far from fit-for-purpose. Roberts and Smith (2014) suggest that prevalent methods of initial and diagnostic assessment across the FAVE sector are not fit for purpose and do little to inform student-centric pedagogy. Moreover, GCSE English teachers can have significant caseloads of teaching, with some assigned at least five or six GCSE English classes, with up to eighteen or twenty students in each class. This can result in teachers being responsible for over one-hundred GCSE English students in the course of an academic year, a unique and startling figure in its own right. This is compounded by the limited number of available teaching hours (one year to achieve the qualification instead of two).

This brings us to the issue of the assessment practices on these GCSE English programmes. GCSE English comprises two end-point summative assessment exams. Each contributes a 50% weighting to a student's final grade in the subject. Both exams take place at the end of the programme, typically in May or June. In my college, all GCSE English teachers are required to conduct milestone assessments with their students, comprising mock papers that mimic the end-point summative assessment exams, at specific intervals throughout the year. The purpose of these milestone assessments is twofold:

1. To conduct an ***assessment of learning*** and determine how students have performed against the examination specifications. Students receive a mark which can be aligned to a grade, in a manner identical to the end-point summative exam.
2. To conduct an ***assessment for learning***, and provide feedback to the student noting what they have done well and where they might improve in future attempts.

In the 2017-18 academic year two of these milestone assessments are completed by students, one in the first term in late-October and one in the second term in mid-March.

The ideas that became the basis for this study were formed as a result of my participation in the college's 2016-17 academic year's milestone assessment cycle, which followed a similar pattern to the one listed above. From experience, I knew that the assessment of these papers was an incredibly demanding task in terms of

the time investment required, with each paper taking between 30-40 minutes to assess, and some teachers having over one-hundred papers to assess in a short window of time. Moreover, it was apparent that despite the teaching team working closely together in the same room when assessing the papers and attempting to standardise assessment judgements, there were likely disparities in how the assessment criteria was being interpreted across different teachers. It is important to point out that this was not so much the fault of any individual but more of an issue of differing interpretations of the assessment criteria.

One of the guiding principles of formative assessment is that learners need to be able to 'see' what success (in all its diverse forms) 'looks like'. A central tenet of this thesis is that teachers need to be able to 'see' this too. It is also interesting to note how assessment for learning was taking place, what form feedback was taking and how this was being communicated to students. The milestone assessments placed an emphasis on the measurement of student performance that could then be reported in the form of a score and grade, but was this at the expense of effective assessment for learning judgements that could inform students of their next steps in learning? These ideas underpin my research questions for this study. These are discussed in greater length in Chapter Three of this thesis.

In considering the above context, there are some questions that we need to ask. How are summative assessment decisions being arrived at? What of the teachers' judgements that are forming the basis for such assessment decisions? In order to answer these questions, we need to consider the role of the teacher in arriving at assessment judgements, and indeed the process of assessment as one of forming a

judgement. It is fundamental that we address this at this point, as a teacher's ability to equitably and accurately judge a student's knowledge of something requires not only an in-depth knowledge of the subject, but also an awareness of how their judgement aligns with that of other assessors.

Before exploring the above questions and the research problem in further depth, it is important to acknowledge that this enquiry is a form of practice-focused research. This is central to the research context. The issues that are presented above, and that are discussed in more depth in remainder of Chapter One, are intended to be an authentic representation of those that GCSE English practitioners face when teaching and assessing in Further Education settings. My intention is to convey these issues and experiences as accurately as possible and with sensitivity in regard to the context in which genuine practitioners are located and in which they encounter their own everyday experiences of practice. My aim here is to ensure that the subsequent findings and discussions in the thesis are trustworthy, authentic, meaningful and that they resonate with the experiences of practitioners working in similar contexts. It is also important to acknowledge my own positionality in the thesis. I am writing from the perspective of a practitioner, rather than an external agent. I am writing as in 'insider' in the research. This influences a number of factors in the focus of the research and in its design. The impetus for this enquiry is driven by my professional interest in addressing issues in assessment theory and practice and is a reflection on challenges that practitioners (including myself) face when teaching GCSE English. I hope that the experiences of all the practitioners involved in this study are given voice in the research and that this will help to enrich the discussions that follow.

## The Problem

### Critical incident – August 2017 review of mock vs actual grades

The idea for this research, and the problem and issues it seeks to investigate and explore, initially came about as a result of data trends that became apparent when reviewing the previous academic year's (2016-17) programme. During the 2016-17 academic year, all GCSE English students at the college completed a mock assessment in February 2017 as a precursor to their final exam. The original intention here was to identify future learning needs and determine their performance in an exam style scenario. When these mock results were compared with the final GCSE grades that were released in August 2017 some interesting trends in performance became evident. The table below features mock grade and actual grade data from a sample of 127 students across both 16-18 and adult programmes. For reference, the mock exam is scored out of a total of 80 marks and each grade boundary spans approximately five marks. A discrepancy of two grades would suggest a student has therefore seen a ten-mark swing in their performance.

| Change in grade performance between mock and final exam | Number of students |
|---|---|
| Increase by 3 whole grades | 4 |
| Increase by 2 whole grades | 18 |
| Increase by 1 whole grade | 36 |
| Remained the same | 27 |
| Decrease by 1 whole grade | 21 |
| Decrease by 2 whole grades | 13 |
| Decrease by 3 whole grades | 7 |

*(Figure 1.1) - mock performance vs. final grade in 127 students in the 2016-17 academic year. Full table available in appendix (appendix item 8.3)*

Apparent in this data set are significant differences between the predicted and actual performance of students between their mock exam in February 2017 and actual their exam in June 2017. We can observe that 58 students improved on their predicted grade, with some improving by two and even three grades. We can also note that 41 students saw their performance decrease, again in some instances by two or three grades. The increased performance in mock vs. actual grades might be accounted for by the fact that mock exams are preliminary exercises and students might not be as prepared or motivated as with actual exams. Or perhaps feedback from their mock performance helped them improve their grade. The same cannot be said for the decreased performance, however. When taken as a whole, this data set hints at possible examples of both over- and undermarking. In parallel with this, we can assume that formative feedback that came about following a quantified judgement of performance might itself be inappropriate and misaligned to a student's actual ability level. Questions of interest here are as follows: how feasible is it for teachers to assess work in both summative and formative modes accurately and effectively under the conditions that are imposed in mock exams? Are there shortcomings in teacher judgement and expertise that lead to the above disparities in results? A first step in exploring these questions further is to consider the practice of assessment (any form of assessment) as involving a judgement, and how the subject context impacts on the judgement process.

**Assessment as a judgement**

Judgement is fundamental to the assessment process, to the extent that the two terms could be considered synonymous with one another. In ideal circumstances the person making a judgement in any assessment scenario will be knowledgeable and experienced in their discipline, have a comprehensive understanding of the standards to which they are assessing and arrive at decisions based on tangible evidence. Each one of the aforementioned requirements is a factor in what effective teachers should strive to already hold and maintain when arriving at judgements. However, in view of these requirements, questions persist: how might assessment standards prevent or facilitate the forming of a judgement? What might constitute good tangible evidence - and what of evidence that does not align with standards? Considering the answers to these questions is crucial if we are to unpack what judgement in educational assessment contexts entails and how it can be employed to positively impact on student learning.

Firstly, let us draw a distinction between judgement as a form of measurement and judgement as a practice. Joughin (2008) observes that 'Assessment as judging achievement draws attention to the nature of assessment as the exercise of professional judgement, standing in contrast to misplaced notions of assessment as measurement' (2008:3). The metaphorical entailments accompanying the idea of 'measuring' learning seem to indicate that learning exists in a tangible form, and that as such its size or amount can be ascertained by use of an instrument or device. Such ideas can perhaps be aligned with what Sfard (1998) terms the acquisition metaphor of learning in which learning is understood to be something through which individuals gain possession of knowledge. The language associated with this

tradition ‘makes us think about the human mind as a container to be filled with certain materials and about the learner as becoming an owner of these materials’ (1998:5). Within the concept of learning as a measurable entity, there is a dependency on the instrument being used for measurement, namely the assessment tasks that students are completing and its respective mark scheme, to explicitly state the learning outcomes for the benefit of the assessor.

This can be problematic. Knight (2007) notes that many learning outcomes not only defy measurement but are extraordinarily difficult to judge. Dunne’s (1993) critique of the technical-rationalist approach to assessment is that such models ‘seemed to arise from the exclusiveness of its concern with instructional outcomes and its corresponding neglect of teaching as an engagement or a process, as well as its inattention to the experiential dimension of learning’ (1993:5). Nonetheless, such instances are commonplace. Ambiguity found in the language of learning outcomes leaves them open to interpretation by teachers who are required to negotiate this subjectivity, often in relative isolation from one another.

To exemplify this point, the image below (figure 1.2) is an excerpt taken from a portion of an AQA GCSE English mock exam paper mark scheme. This mark scheme is used by teachers to assess the creative writing response that students complete as part of their milestone assessment at the college, and in their final summative endpoint exam at the end of the programme. The image features descriptors relating to the candidate’s proficiency in their use of content and organisational features at an approximate grade 4 Level (a grade ‘C’ in pre-reform equivalency).

| | | | |
|---|---|---|---|
| Level 2<br><br>7-12 marks<br><br>**Some success** | Upper Level 2<br><br>10-12 marks | **Content**<br>• Some sustained attempt to match register to audience<br>• Some sustained attempt to match purpose<br>• Conscious use of vocabulary with some use of linguistic devices<br><br>**Organisation**<br>• Some use of structural features<br>• Increasing variety of linked and relevant ideas<br>• Some use of paragraphs and some use of discourse markers | At the top of the range, a student's response will meet all of the skills descriptors for Content and Organisation<br><br>At the bottom of the range, a student will have the lower range of Level 2 and at least one of the skills descriptors for Content and Organisation from the upper range of Level 2 |

*(Figure 1.2) - excerpt of AQA GCSE English Language paper 1 mark scheme*

In this excerpt we see six standards, three relating to content and three to organisation, in which the student needs to have demonstrated proficiency if they are to be considered to be working at this level. There is also a variable range of marks available for the candidate if they are deemed to have shown proficiency in some but not all of the standards. The language present is highly interpretative (*some sustained attempt…, some use of..., increasing variety...)*, and we can perhaps attribute the subjective nature of the assessment as being responsible for the ambiguity in language evident in this example. It would be near impossible for any mark scheme to specifically prescribe what form a creative writing piece that demonstrates *some success* should resemble. Nevertheless, it is on the interpretation of these descriptors that a teacher's assessment judgement hinges, and this is where we can locate judgement as a practice, rather than an instrument or measure.

For a teacher to arrive at an accurate assessment judgement when assessing a student's creative writing response, their decision is reliant on their experience, skills and knowledge. Teachers may talk of 'getting impressions' and 'gut feelings' when assessing student work that defy articulation through written outcomes and standards (Bell & Cowie, 2001), which seem to suggest judgement to be a tacit process. Moreover, it might be argued that conceptualising judgement as a practice invites the notion that this is something that can be honed and improved through purposeful repetition by the teacher. These same principles cannot be applied to the idea of judgement as a form of measurement, as in this tradition this process is inhibited by only what can be articulated through the assessment instrument and its respective mark scheme. Dunne (1993) notes that 'atomistic objectives may seem worthwhile, however, only if they aggregate over time into qualities of mind and character, such as an ability for independent thought and reflection, a habit for truthfulness, a sense of justice, a care for clarity and expressiveness in writing and speech' (1993:6). For Dunne, the language of learning outcomes and assessment criteria is 'designed precisely to exclude these qualities as being too vague or too open to divergent interpretations' (ibid:6). As a result, when students are required to demonstrate creativity, ingenuity and original thought in learning scenarios, teachers must synthesise their experience, skills and knowledge to work in complementary ways with assessment criteria, and recognise that such criteria are heuristic devices rather than precise and clearly defined instruments of measurement.

In spite of the above, it must be noted that the aforementioned GCSE English Language assessment and others akin to it might be interpreted as being exercises in measuring learning and quantifying student achievement. Within the context of the

milestone assessments this risk is particularly pronounced as its dual purpose as an assessment both *of* and *for* learning may be misinterpreted owing to the need for individual teachers to share the marks and grades with agents other than the student, namely the college's hierarchy. In such circumstances, the measurement takes precedent. This 'misplaced notion' as Joughin (2008) notes, can perhaps be traced back to the previous discussions on the mischaracterisation of summative assessment solely as an instrumentalised or mechanical tool for assessment, rather than a process that can enable learning. It is important to be aware of the possible perceptions that a teacher may have, whether they are overt in what they say or covert in their actions. This is something that this study will attempt to address.

**Judgement as a practice**

Dunne (1993) maintains that practice is:

> 'A coherent and invariably quite complex set of activities and tasks that has evolved co-operatively and cumulatively over time. It is alive in the community who are its insiders (i.e. its genuine practitioners) and it stays alive only so long as they sustain a commitment to creatively develop and extend it – sometimes by shifts which may at the time seem dramatic or even subversive. Central to any such practice are standards of excellence, themselves subject to development and redefinition, which demand responsiveness from those who are, or are trying to become practitioners'
>
> (Dunne, 2005:152-153)

In accepting that assessment manifests itself as a result of judgement practice, we must appreciate the need for teachers to have, as Boud (2007) suggests, 'the capacity to evaluate evidence, appraise situations and circumstances astutely, to draw sound conclusions and act in accordance with this analysis' (2007:1). The distillation of these composite elements of a judgement align with Aristotle's (2011) concept of *phronesis*, commonly translated to be a kind of practical wisdom, a key component of which is the ability to deliberate on a problem. It is advanced that deliberation is a 'sort of investigating' and good deliberation involves investigating what to do for an end result that is fair and just (2011:126). Dunne (1993), in a similar interpretation offers the idea that phronesis is the disposition towards perceptiveness, describing it as a:

> '…disposition for perceiving, or having insight…it helps one mediate between more generic, habitual knowledge and the particularities of any given action-situation, and it involves perceptiveness in so far as its apprehensions are not deductively derived, but are freshly generated in response to the particularity of this situation and the individual's involvement in it now' (1993:51)

As we appreciate the complexity of judgement as a form of practice, we too must acknowledge that if accurate and fair judgements are arrived at through *phronesis*, tacit knowledge and rich, varied experiences, teachers require time and opportunities to foster the development of these qualities.

The term practice is grounded in the actual application of an idea, belief or method rather than theories that relate to it, although we can also recognise the relation held

between practice as a process, and practice as an act of rehearsing a behaviour with the intent to improve or master it. These two concepts are far from disparate from one another. Effective judgement practice is not something that we can assume all teachers have, nor is it an easy thing to assess or measure if we are to determine if teachers are capable of forming competent judgements, as by definition the successful application of this practice is contingent on the teacher having command of the idea, belief or method through which the judgement is being formed.

So where can we locate judgement practice? Are the ideas, beliefs and methods that comprise judgement practice constructed and maintained by an individual, or are they established and sustained by communities of practitioners that each contribute to what the base norms for these are? It might be that both are correct. Filer (2002) argues that assessment practice is a 'social practice' and yields 'social products', noting that:

> 'Its wider functions are concerned with social differentiation and reproduction, social control and the legitimizing of particular forms of knowledge and culture of socially powerful group, [...] and so the social and cultural values, perceptions, interpretations and power relations of assessors and assessed carry important implications for processes and outcomes' (Filer, 2002:2).

Filer acknowledges that judgement practice can be located as a social construct that cannot be separated from external influences, whilst also observing that individual interpretation, agency and autonomy in these practices can have significant impact on the equity and validity of the judgements a teacher may form. This chimes closely

with considerations of standards, which are discussed a little later in this chapter. In terms of locating judgement as a practice, we can observe that in the first instance it is a socially constructed concept that is then adopted and interpreted by the individual.

The problem that arises from this line of thinking is that teachers need opportunities to become conversant in judgement practices, but in order to do so they are dependent on others to initially induct them, and then support them in maintaining currency, before they themselves eventually become the co-owners of these practices. This Chapter has already outlined the scale of assessment that teachers at my college face. Some have over one-hundred students they are responsible for, with each student's script requiring approximately 30-40 minutes of assessment time for an effective judgement to be formed. In view of this, and appreciating that the majority of these teachers that are relatively new to assessing the qualification, the opportunities to access and be informed by the approaches, norms and knowledge that make up judgement practice as sustained by the wider GCSE English teaching community are inevitably limited. The wider implications of this might be that students are not receiving accurate feedback that informs them, themselves a likely novice in the interpretation of assessment standards, what they have done well or where they can improve. Much of the discussion in Chapter Two is centred on exploring further the relationship held between judgement and practice. This goes on to underpin the methodological approach explained and justified in Chapter Three.

**Assessment standards**

This Chapter has so far established that judgement is a form of practice, and that effective judgements are conditional on a teacher's ability to negotiate different and sometimes challenging assessment contexts by drawing on their experience, skills and knowledge. Additionally, these discussions have recognised that phronesis and tacit knowledge also play a crucial part in how these judgements are formed. This Chapter also notes that such knowledge is difficult to codify in mark schemes or assessment criteria. In this event, the raises questions of the extent to which is it right to assume that assessment standards based on assessment criteria in some way inhibitive or restrictive? Why is it that we have standards and assessment criterial in any event - if teachers are conversant with the multiple facets that make up their subject should it not be right that they are the gatekeepers of standard rather than lists of centrally devised and highly prescribed lists of standards?

Let us first define what is meant by the term 'standard'. Within an educational context, standards are understood to be 'fixed points of reference for assessing individual students' (Sadler, 1987:191). Various types of standards can include:

- standards as moral or ethical imperatives (what someone *should* do)
- standards as legal or regulatory requirements (what someone *must* do)
- standards as target benchmarks (expected practice or performance)
- standards as arbiters of quality (relative success or merit)
- standards as milestones (progressive or developmental targets)

(Adapted from Maxwell, 2001)

It is important to recognise the extraneous influence that some of these types of standards hold within educational circles, and in particular standards centred on

target benchmarks. On the international stage many standards serve as target benchmarks that carry an expectation on a designated level of practice or performance from students in relation to specific domains, as seen in the Programme for International Student Assessment (PISA) rankings that compare various nations' scholastic performance in mathematics, science and reading. The aim of the PISA rankings is to enable governments to develop and improve the educational policies, which once developed, are themselves aligned to standards which then act as target benchmarks. As the effects of such policies and their commitment to tangible outcomes filter down to the classroom there are inevitable consequences for teachers, and much has been written about the unintended consequences of high-stakes accountability policies and how they can undermine quality teaching and learning, and equity related efforts (Nichols and Berliner, 2007; Stobart, 2008; Klendowski and Wyatt-Smith, 2013). Whilst this research study does not seek to foreground the resulting effects of national policy-aligned standards, their influence will punctuate much of the discussion that follows.

We can locate our interest in the last two types of standard that are noted above, as they neatly align with the key aspects of summative and formative assessment respectively, namely, assessing the quality of an educational product (relative success or merit) and assessing development or improvement over time (progressive or developmental targets). It is important to recognise that teachers, by their very craft, are members of a community of practitioners that stretches far beyond their own immediate horizons. Through their teaching of a subject they automatically gain membership to this community, which carries with its own significant implications for any teaching professional. They alone are not

gatekeepers of their discipline, and as a result they cannot act as so when teaching, assessing or otherwise. This carries consequences when forming a judgement too, as Klendowski and Wyatt-Smith (2013) note:

> 'Judgment is inherently a private practice: the actual influences on and bases for judgement typically remain private. It is only when standards are defined and applied in standards-referenced judgement practice that standards can become published indexes or features of quality against which judgement can be made available for scrutiny and, thereby, made defensible' (2013:13).

The argument here raises an important matter regarding the defensibility of a judgement. If judgements are formed without reference to external sources it becomes very difficult to verify the validity and credibility of the judgement formed. A teacher's understanding of creative writing, for example, might be far removed from a colleague's on account of the rich differences the two professionals hold in skills, experience and knowledge. Resultantly, the two judgements reached if formed solely of their own knowing of the topic are likely to deviate to such a degree that the conclusions drawn might be unrecognisable from one another. Applying the principles of this scenario across a whole cohort of students and teachers, it is possible to see where issues of equity, fairness and pedagogical worth might arise.

This Chapter has so far sketched a broad picture that seems to indicate the validity of standards in ensuring the maintenance of minimum levels of achievement. It has not yet however sought to compare the necessity of standards with the aforementioned concept of tacit knowledge. Whilst standards can ensure equity,

transparency and conformity across a large number of teaching professionals, they still fall short in being able to define the role of tacit knowledge and determine the extent to which it should help form the judgement being arrived at in context in and across a community of assessors. In this sense, subscribing to technical-rationalist models of assessment practice based upon the measurement of learning outcomes and centrally prescribed assessment criterial can result in what Dunne (1993) warns to be a narrowing of what it is that is being assessed, sometimes at the expense of what he considers the 'significant achievements of education' (1993:6), comprising independent thought, reflection and expressiveness. Looking again at Figure 1.2, we can recognise that there are no explicit instructions for teachers on how to judge to these standards. Rather, the assessment criteria advocate despite their positivist overcoats invite a more open, interpretative and even a holistic approach to judgement than their pretentions toward objectivity might suggest. Despite its omission on account of its fluidity, variability and inability to define, tacit knowledge and an ability to be deliberative and perceptive is relied upon in the use of assessment criteria.

**The subject context for this enquiry: creative writing**

At this juncture we can also locate discussions centred on the challenges arising when teachers attempt to form a judgement on an artefact of work that does not feature in the assessment standards, either through explicit reference or through tacit acknowledgement. It is here that we locate the reasons why enquiry focuses on creative writing as the subject context. Creative writing is a pursuit which is hard to define and articulate solely through descriptors and standards. In capturing what

creative writing entails, Morley (2007) notes ‘think of an empty page open space. It possesses no dimension; human time makes no claim. Everything is possible...Anything can grow in it...there is no constraint, except the honesty of the writer and the scope of imagination.’ (2007:1). In view of this, questions might arise asking what makes for ‘good’ evidence in creative writing? It is the breadth of possibilities that stem from creative writing as a discipline with the subject of the English Language that make it such an interesting area of focus in this enquiry. Morley’s (2007) assertion that ‘everything is possible’ in creative writing is of course true, but we can appreciate that it is possible to discern a good piece of creative writing from a bad one. Specifically, for a teacher-assessor, this ability to know what constitutes ‘good’ creative writing is critical. This thesis sets out to explore this issue in depth.

In assessment scenarios such as examinations, there is a degree of conformity required of students when composing a piece of creative writing, in that there is a need to respond to a specific assessment brief. Beyond a conformity to topic and length of the piece, students are then afforded the breadth of their own imagination to express themselves in whatever manner they wish. When forming judgements on student artefacts it is necessary for teachers to appreciate the breadth of the subject that might be employed by students, including elements that might not feature on the curriculum specificiation. This is of particular importance with regard to the principles and formative assessment practices in a pedagogical approach which has come to be known as *Assessment for Learning*. The deeper point here is that although students might be exhibiting skills and knowledge that do not map directly on to the prescribed curriculum content (this is in no way a detriment to their learning), they

are perhaps demonstrating skills and knowledge that can be mapped to standards in more complex and subtle ways . However, such instances require teachers to be conversant with this breadth interpretation of curriculum and assessment standards. Teachers need to be open to different ideas, concepts, ways of working, ways of interpreting, and seeing in order to recognise and capture actual student achievement in the subject of creative writing. This is a central aspect of the thesis. A broader discussion of this features in Chapter Two.

Ultimately, we can appreciate that standards play an important role in the maintenance of integrity and content of qualifications and subjects but it is the teacher who interprets these standards in a given context, and in doing so is required to draw upon their own skills, experience and knowledge to form a judgement. Questions still remain on these matters, including, what is tacit knowledge? Is it something that an individual can hold? Does it exist as a socially constructed concept within teaching communities? Or are both true, in that it moves between the two? The following chapters of this thesis attempt to explore answers to these. Further discussions feature in Chapter Two which elaborates on the role of assessment standards in making a judgement, and how teaching professionals must negotiate between tangible standards and their own tacit understanding and knowledge within a subject discipline to arrive at a judgement. Attention is paid too to how standards act as arbiters of quality and progress, whilst ensuring that equity and minimum expectations are met, how this can be inhibitive, and what prevailing research in this field suggests can be done to traverse this.

## Concluding remarks

This Chapter set out to initially establish the differences and similarities in formative and summative assessment practice, and recognise how summative assessment has come to be misunderstood as a rudimentary measure of learning, rather than a practice of judgement that can help contribute towards the acculturation of learning opportunities if applied in specific contexts. It has chartered the significant shifts in policy that English as a subject in the FAVE sector has endured in recent years and seen how principles of formative and summative assessment practice are applied within a specific context in a large FE college. Central to this are the teachers of this specialist subject, who operate in challenging conditions and negotiate between an adherence to institutional accountability mechanisms that require quantified measures of learning and the development of pedagogies of contingencies for their students when assessing their work.

It is against this backdrop that we can locate the issue that this research seeks to address. With teachers facing multiple agendas when assessing student work, there are genuine risks that educationally valuable assessment practice, that is to say forming a fully realised summative judgement that also provides an assessment *for* learning, will be traded off in favour of a solely reductive measurement of performance. The latter half of the chapter has put forward a case for understanding assessment judgement to be recognised as a social practice that exists within a community of practitioners, rather than in isolation with individual teachers, and how such domains, through their establishment and maintenance of a shared tacit

understanding of such practices, can help teachers better understand assessment standards and reach astute, equitable and educationally valuable judgements which rigidly structured systems of assessment criteria struggle to do.

# Chapter 2 - Literature Review

## Introduction to the chapter

Chapter One made brief forays into discussions of some of the predominant issues in assessment theory and practice that his thesis sets out to explore. These include considerations of questions such as, what it means for assessment to be an educationally valuable practice; the challenges of judging student performance in a specific subject and context and the complex and dynamic relationship that exists between the act of forming a judgement and understanding judgment in assessment contexts as a practice. This Chapter explores these matters in further detail and in some depth. It looks to theoretical and historical perspectives to foreground some of the main contributions to this discourse from the literature. In addition, much attention will be paid to prior empirical research that has been undertaken, in both similar and varying contexts to the one in which this enquiry is situated. The purpose here is to provide frames of reference which may be helpful in informing this enquiry.

This Chapter is presented in two sections. The first leads a discussion on different understandings of skill development and varying forms of knowledge. Consideration is also given to how teachers can develop skills and knowledge in the field of assessment theory and practice, and what benefit these skills and this knowledge serve in different contexts. The second section of this Chapter explores the shifting landscapes of English within the FAVE sector in recent years and the domain of creative writing that forms the focus of this study. The same section also introduces and discusses the concept of Adaptive Comparative Judgement (ACJ), on which this thesis is based. Chapter Three then presents a justification for the methodology and

methods employed in the study in relation to the research question.

## The development of 'skill' in judgement

We have so far established that judgement requires a comprehensive and nuanced understanding of a discipline in order for it to be reasonable, reliable and accurate. What we have not yet explored is the nature of judgement as a skill and form of practical knowledge acquired through experience and honed through trained practice. As yet, there exists no procedure or 'good practice' to follow when forming an assessment judgement in GCSE creative writing assessment tasks, and it remains a hugely interpretive and individualistic exercise, albeit one that is intrinsically attached to social contexts. As such, considerations into how the development of one's skill in making practical judgements might be fostered so it can successfully operate in varying and complex contexts are important.

In *The Craftsman,* Sennett (2008) ventures the notion that a skill is 'a trained practice'. He argues that skill development depends on how repetition is organized and that as a person develops skill, the contents of what he or she repeats change. Skill development is attained, Sennett argues, through careful observation, imitation and repetition, and progresses in line with practice that leads to valuable experience. The nature of this practice he argues must be 'purposeful', and not solely observational or imitational as one develops an increasing competency in the skill. References are drawn to the seminal paper produced by Ericsson et al. (1993) into the role of *Deliberate Practice in the Acquisition of Expert Performance*, which defines deliberate practice as 'those activities that have been found most effective in

improving performance[...] it is a highly structured activity [...] requires effort and is not inherently enjoyable' (1993:367-368). It concludes that to attain equivalency with elite performers in a discipline an equivalent of a minimum of 10 years of intense practice is necessary, something that Sennett compares to the attaining master status within a craft.

From the above points of view, a skill is something that can be acquired and honed regardless of any apparent innate talent, and can be so through optimal means; the crucial elements being that that the individual is learning within their zone of proximal development (ZPD) (Vygotsky, 1978; Ericsson et al., 1993), that they receive immediate informative feedback on their performance and they should repeatedly perform the same or similar tasks (Ericsson et al., 1993). Sennett (2008) proposes that a person who cannot observe, cannot enter into an open dialogue with someone more skilled about how to improve a practice. This is because skill development is reliant on problem-finding, problem-solving and critique and these help to challenge notions of understanding, and this cannot take place in solitary or one-off events alone. Rather, he contends conversation through dialogue encourages individuals to develop their own interpretations and applications of new knowledge and these are fundamental in progressing a skill so that it becomes 'more problem attuned'. In exploring Aristotle's (2011) dimensions of phronesis, Chinn, Maeve, and Bostwick (1997) suggest that the perceptiveness that one can become trained in implies an aesthetic knowing or 'connoisseurship—a keenly trained 'eye' and 'ear' and a 'feel' for a practice (1997:85).

Sennett maintains that as a skill is developed, technique is no longer a mechanical activity and that people can feel fully and think deeply about what they are doing once they do it well. Individuals become cognizant of their aptitude in a skill, and can draw on both their explicit and tacit understanding of a concept to realise high standards of work. This aligns with his conception of the development of skill that forms a capacity to both problem-solve and problem-find as it grows. This is often a necessity as a result of the constantly evolving nature of the contexts in which skilled workers operate. Sennett suggests that problem- finding and problem-solving exist in an 'experimental rhythm' with one another, and that a skilled individual will have the foresight to recognise problems before they might occur, in addition to solving problems when they might arise. We can perhaps align these ideas with the skill necessary for assessment judgements of student scripts. A lesser skilled assessor might fail to recognise the opportunities to assess *for* learning when marking a script in a summative procedure as a result of their subconscious subscription to the forming of a judgement aligned to a measurement paradigm. This might be contrasted with a higher skilled assessor who can recognise the need to form assessments both *of* and *for* learning, so that they can form judgements that take stock of what has been achieved, and what opportunities exist for future development. The latter example here represents Sennett's idea of problem-finding, through which a skilled individual can negotiate evolving circumstances and still maintain high standards of quality.

## Standards and quality within skill development

The notion of standards and quality is one that is closely aligned to the development of a skill. Sennett (2008) invites us to consider what we mean by 'good work'. He observes that:

> 'Often we subscribe to a standard of correctness that is rarely if ever reached. We might alternatively work according to the standard of what is possible, just good enough—but this can also be a recipe for frustration. The desire to do good work is seldom satisfied by just getting by.' (2008: 45)

For Sennett, standards can act as inhibitors to highly skilled individuals but serve to ensure the maintenance of quality markers. The risk in adhering solely to such standards is that they result in a loss of genuine craft. Whilst highly skilled practitioners operate in community oriented socially structured groups, standards are often imposed by bodies external to these individuals that remain unreactive to change until goals, procedures and desired results for a policy have been mapped in advance, and neglect dialogic and collaborative ways of working. Sennett terms these 'closed-knowledge systems'. He maintains that at the higher levels of skill, there is a 'constant interplay between tacit knowledge and self-conscious awareness, the tacit knowledge serving as an anchor, the explicit awareness serving as a critique and corrective' (2008:50). The problem with this, stems from the long-held concerns of proponents of absolutist standards of quality, who view the amalgam of tacit and explicit knowledge as an experiential standard lacking specification and form. Sennett differentiates between conflicting measures of quality 'one based on correctness, the other on practical experience' (2008:52) and notes

that they cannot be reconciled with one another; in such instances the diverging claims of tacit and explicit knowledge pull the skilled practitioner in contrary directions.

In order to find a purposeful direction between tacit and explicit knowledge, practitioners must negotiate between their explicit and tacit knowledge, in what Sennett terms ‘liminal spaces’. Sennett cites the example of medical practitioners who operate within parameters aligned to a Fordist model of medicine, in which there must be a specific illness to treat, and that evaluation of a doctor’s performance will be made by comparing the time spent treating a patient with how many patients get well. In such systems these imposed standards restrict the practitioner to a classifying model that often fails to address the problem. Sennett notes that ‘reality doesn’t fit well inside this classifying model, and [...] good treatment has to admit experiment’ (2008:49). It is here that the importance and power of liminal spaces that can be found. Sennett notes that through dialogue with a patient, medical practitioners operate in a liminal zone between problem-solving and problem-finding, and can glean clues about ailments that might escape a diagnostic checklist. This medical analogy lends itself to educational contexts quite neatly when mapped to skilled judgements of student performance. There is a recognition in the highly skilled individual that standards and quality can infringe on the establishing of adequate judgements, and that one’s own experience and tacit knowledge can be employed to work in tandem with these standards. This, Sennett maintains, is a conscious undertaking; ‘bedded in too comfortably, people will neglect the higher standard; it is by arousing self-consciousness that the worker is driven to do better’ (2008:51).

In view of the above, it might be fair to conclude that standards are inherently flawed and forsake genuine innovation and excellence in practice in favour of ensuring a minimum level of performance. This is not necessarily the case. Standards and quality are crucial in defining what a competent skilled performance looks like and what it comprises. Sennett's citing of medical practitioners working against the bureaucracy imposed on them is a helpful example of how standards and quality markers can have unintended consequences when they counterintuitively work against practitioners when they intended and out in place to do good. Even so, one can appreciate the need for standards and the defining of what 'good' medical practice looks like in a discipline as wide-reaching and vital as national health care. The same can be said of education. A lack of standards will likely lead to inconsistencies in practice and potential mediocrity and negligence. In view of this, the question that faces us is not a question of the extent to which should skills should be held accountable to standards and quality markers, but rather how many standards are appropriate for this particular skill and how much flexibility should they offer, and what are viable quality markers in consideration of the context?

## Tacit knowledge and professional learning

This thesis charts some of the challenges facing GCSE English teachers in FAVE contexts. For example, recent sweeping changes to the curriculum landscape, the difficulty in aligning effective pedagogies and approaches to assessment to the realities of institutional contexts and the necessity to form effective judgements on student performance that assess learning both in and after the event are but a few of the issues at work here. This next section of this Chapter briefly explores the concept

of professional development for FAVE-based English teachers. It is argued that the aforementioned issues and challenges that teachers face can each be overcome through effective professional development, and this is what makes this a topic worthy of exploration. The topic of professional development is a vast one, and as such this discussion will focus on establishing how the previously mentioned concept of tacit knowledge is cultivated in teachers, as this is a critical facet in the forming of an effective professional judgement.

We can perhaps logically begin here with an exploration of what is understood by the term tacit knowledge. Polanyi (1966) first conceptualised tacit knowledge as relating to both perception and scientific thinking, proposing that 'we know more than we can tell' (1966:4). The nature of tacit knowledge is stressed to be largely experiential, through which it can be passed on by demonstration, example and practice (Elliot et al., 2011). Prominent schools of thought have conceptualised it according to three main features: firstly, that it is acquired without a high degree of direct input of others, but rather from an individual's experience of operating a specific context; secondly, that it is procedural in nature and concerns how best to undertake specific tasks in certain situations; thirdly, that how an individual uses tacit knowledge is intrinsically bound to their own circumstances, disposition and personality and may lead them to take actions that are effective in serving their own personal goals and agendas (Sternberg, 1995; Sternberg, 1997; Sternberg & Hedlund, 2002; Grigorenko et al., 2006; Elliot et al. 2011).

We can note a difference between tacit knowledge and explicit knowledge, sometimes termed codified knowledge. It has been argued that both types of

knowledge are necessary in the establishing of an accurate judgement. So what relationship is held between these two banks of knowledge? Research conducted by Greenhalgh et al. (2008) explores how multidisciplinary teams of medical practitioners balanced encoded knowledge in the form of standardised outcome measurement with tacit knowledge, in the form of intuitive judgement, clinical experience and expertise. What their analysis suggests is that clinicians draw on tacit knowledge to supplement, adjust or dismiss 'the scores', derived from a standardised assessment scale, in making judgements about a patient's likely progress in rehabilitation. They conclude by noting that standardised outcome measures can 'support, rather than determine clinical judgement' and that 'tacit knowledge is essential to produce and interpret this form of encoded knowledge and to balance its significance against other information about the patient' (2008:1) The findings here, although derived from a discipline other than education, point to recognition from a body of professionals that tacit knowledge can supplant codified, standardised knowledge in the forming of a judgement. There is also a suggestion here that tacit knowledge plays a crucial part in the interpretation of codified knowledge, be that the interpretation of assessment standards or otherwise.

We can trace commonalities regarding the development of tacit knowledge, namely that it is intrinsically bound to active participation in an activity or pursuit and that it is highly subjective owing to its individual interpretative nature and as such is difficult to prescribe or standardise. Polanyi (1958, 1966) presents the example of using a hammer and how the individual sensory feedback we receive can lead to the individual using the tool more skilfully, through experience and conscious situatedness. This principle is sustained when the skill being applied is one where

self-feedback is not possible. Schools of thought sustain that social interaction and collaboration, through which demonstration, observation and feedback can occur, play an important role in the establishment of tacit knowledge, arguing that it is through interaction with others that tacit knowledge can be broadened as individuals are presented with new, unfolding and challenging contexts to which they must apply current knowledge. In line with this, Fox (2000) highlights the problem of viewing tacit knowledge as residing in individuals and, with reference to Lave and Wenger (1991), suggests that it should be seen as part of a community of practice.

In a study exploring the differences between experienced teachers that have tacit knowledge and novice teachers that do not, Elliot et al. (2011) identify that:

> ‘experienced teachers and novices do not differ significantly in terms of the capacity to identify good solutions to situational problems, but rather, they differ significantly in their skills at identifying poor solutions to these same problems [...] This suggests that tacit knowledge in this particular domain is not so much a matter of learning how best to approach a problem so much as it is about learning how to avoid making a really bad decision’ (2011:98).

This matter of knowing what to do when you do not have all the required information or expertise available to you is a pertinent one for this study, as the conclusion reached here by Elliot et al. suggests that a deficit of tacit knowledge can negatively impact teachers' judgements and decision making. In scenarios where teachers might be unsure of a judgement decision they are required to make, tacit knowledge can provide insight into what a favourable course of action might be in view of the

circumstances, on account of drawing from experience and prior encounters where similar, but not identical, challenges were resolved.

But tacit knowledge is more than just having experience in a specific domain. It is what is developed and honed as a result of experience. Whilst explicit knowledge can be codified, in forms similar to those seen in assessment mark schemes, tacit knowledge remains impossible to define in written form. Despite this, it does exist in internal cognitive forms, procedures and logic as would be found in explicit knowledge, but through individualistic interpretation and cognitive mediation these are synthesised both consciously and subconsciously to form meaning. For now, we can conclude that tacit knowledge and experience play a significant role in the act of forming a judgement and that considering how such knowledge can be developed in professional settings is important in the development of assessment theory and practice.

## Professional learning and opportunities for GCSE English teachers in the FAVE sector

It is with this in mind that we can turn our attention to research and activities centred on continuing professional development (CPD) as a means of fostering commonality in understanding across teachers. Broad (2015, 2016) observes that CPD activities can enable teachers to 'network with, collect and transport tacit knowledge that has already been semi-codified for other purposes [...] within the specialist area.' The role that collaboration with colleagues plays in this process is acknowledged here, 'it

is not, as with codified knowledge, found in textbooks and curricula documentation, and is not produced for the consumption of students' (2015:9). Broad (2016) proposes that this 'semi-codified' knowledge lies between explicit and tacit knowledge in forms such as artefacts, processes and mechanisms that are used to capture workplace knowledge but that are not organised specifically for curricula purposes. Zollo and Winter (2002), in their study of Japanese corporations conceptualise semi-codified knowledge in a similar manner, describing it as diffused, fuzzy, institutionally based knowledge that is only available to trustworthy insiders.

In this first instance, we consider the role of national CPD initiatives that have impacted on the sector. In recognising the national need for opportunities to upskill English teachers, following the *Maths and English provision in post-16 education (2014)* report, The Education and Training Foundation developed an 'English Pipeline' comprising a series of workshops, webinars and comprehensive Level 5 subject specialist modules tailored to the teaching of GCSE English in FAVE settings. By the start of the 2015/16 academic year, 4,000 Further Education English teachers had participated in at least one of these English enhancement programmes, (ETF, 2015) demonstrating the deep desire and concerted effort that both teachers and their institutions exhibited to best prepare for the changing landscape. At the time of writing, this number now stands at over 11,000 teachers (ETF, 2018).

These programmes are designed with the specific intentions to encourage cross-peer and cross-institution collaboration, discussion and problem solving in group settings, and echo Lave and Wenger's (1991) concepts of communities of practice.

Despite the clear progress that has been made through national initiatives such as the 'English Pipeline', there are still shortcomings with viewing this initiative alone as a panacea to the wider issues of continuing professional development facing English teachers based in the sector. Many of the activities such as those in the 'English Pipeline' comprise standalone workshops that remove teachers from their own context so as to place them in artificial communities that exist only for finite periods, often only a day or less. As a result, teachers are not able to tackle genuinely challenging authentic issues that they face in their practice day to day, and accordingly apply existing knowledge to new contexts leading to the forming of new tacit knowledge. The development of Broad's conception of 'semi-codified' knowledge experiences a diminished role in these activities too as a result of the same circumstances. In instances where teachers participate in professional development activities as events outside of their institutions the processes and mechanisms that their institution already use do not feature. This is not to say that initiatives such as the 'English Pipeline' are not capable of fostering the development of tacit knowledge of teachers, but rather that the optimum conditions for its development exist in the teachers' own immediate professional context.

Collaborations that are fostered between colleagues at the same institution that attempt to facilitate the transmission of tacit subject specialist knowledge between English teach staff members represent rich opportunities to foster growth in all members of the teaching team that participate. Ish-Horowicz (2015) identifies that the trialling of sharing English pedagogy meetings led to 'an improvement in teacher competence and confidence [...], the space to share ideas was appreciated by teachers, and led to more collaborative teaching and planning' (2015:1). Indicated

here is the positive impact that small adjustments to typical working practices within an institution, such as the fostering of collaborative learning communities between teachers, can yield for teachers. What is not apparent here is the tangible impact these activities had on student learning, either as a result of an increase in teacher tacit knowledge of English pedagogy, or otherwise. Whilst this was not a focus of the research, the question remains as to *how* continuing professional development activities can be tailored so as to foster the development of tacit knowledge.

One common activity that teachers undertake with relative frequency is in the standardisation of their assessment decisions. Typically, standardisation sessions held between team members would not be characterised as formal professional development activities, although this might vary depending on the design and running of the session. Standardisation practices vary across contexts, but predominantly comprise two or more teachers seeking to align their assessment judgements with that of the mark scheme and one another. It may also include moderation of other prior assessed work to ascertain the validity and quality of judgements. There are several components to standardisation that align with the tenets of effective CPD for teachers, as outlined above; such activities are grounded within an institution's own quality processes and mechanisms and require the participation of some or all of the teaching staff to contribute towards the completion of the activity. Problem-solving through collaboration and the establishment of common ground understanding across teachers is a goal of standardisation activities. It stands to reason that these activities will help forge tacit knowledge in teachers as they are exposed to other ways of thinking, interpreting and forming a judgement, and perhaps have their own judgements challenged by colleagues.

In spite of the above, the extent to which standardisation practices can be considered to be effective CPD pursuits in which tacit knowledge is forged is another relatively under- researched area. It is hard to draw conclusions from isolated examples of standardisation activities across institutions without an appreciation of the variable methodologies or processes employed in these specific contexts. Some challenges arise from viewing standardisation as valuable CPD regardless of its design or implementation. In instances where standardisation activities are solely oriented to ensuring commonality in the measuring of performance in assessments the nature of the tacit knowledge being transported between teachers might advocate the prosaic interpretation of the assessment standards, rather than a broader appreciation of the text. Much in the same vein, standardising the measurement of performance fulfils an obligation to achieve a common understanding of quantifying performance but does little to establish tacit knowledge in how best to respond to the student in feedback or adopt a pedagogy of contingency. In such instances, these standardisation practices might be considered CPD activities, in that they are developing professionals in their roles and understanding of the subject, but in doing so are perpetuating practices that are pedagogically deficient. This enquiry seeks to cautiously explore how these challenges may be overcome through the methodological approach that is adopted in this study and justified in Chapter Three.

In view of the above, there is still a dearth of understanding relating to the development of tacit pedagogical knowledge in FAVE-based English teachers specifically relating to assessment practices. It remains an under-researched area

within the sector. From professional experience the transmission of tacit and nuanced subject knowledge between colleagues in specific institutions is perhaps the most valuable and impactful and form of professional learning available to them. However more research is necessary in order to fully determine these perceived benefits.

## The Literacy-English divide, and the rise of the National Curriculum

The intention of this section of the chapter is to introduce the shifting landscape on which English curriculum in the FAVE sector is located, and then to broaden the scope to look at perspectives of creative writing as an individual pursuit both inside and outside of context of formalised study spanning the last 60 years or so. In charting this path, it is hoped that we will better understand the challenges and opportunities that exist for English teachers when they judge student performance in creative writing within FAVE contexts.

As noted in Chapter One, the Wolf Report (2011) acts as the catalyst for the adoption of GCSE English for 16-18 year old students in the FAVE sector, which draws its prescribed content from the National Curriculum. This represents a significant shift not least because the predecessor qualifications to GCSEs in the FAVE sector, Basic Skills, Key Skills and Functional Skills, each drew their taught content from the Adult Literacy Core Curriculum standards which were first published in 2001. The use of ‘Literacy’ rather than ‘English’ in the standard’s title bears some

consideration as it provides insight into the ideological aims and intentions of the standards. Some definitions characterise literacy as being associated with comprehension and the ability to act upon information, rather than the production of language through writing (National Adult Literacy Database, n.d; Ofqual, 2012; PISA report, 2012), and it perhaps stands to reason that these are representative of the literacy detailed in the Literacy Core Curriculum standards, as the omission of student writing assessments from on-demand tests in Basic and Key Skills qualifications and subsequent desire for a broader curriculum became one of the catalysts for change following the Wolf report and the subsequent marginalisation of the Literacy Core Curriculum's role in the sector. The marriage of the concept of literacy to FAVE contexts is one that has been forged over some time, and carries with it unshakeable connotations. One such is what Gee (1996) presents as the '…commodity myth' (1996: p. 122), in which:

> 'literacy = functional literacy = skills necessary to function in "today's job market" = market economy = the market = the economy...Literacy is measured out and quantified, like time, work and money [...] We match jobs with "literacy skills" and skills with "economic needs". Literacy, thus, becomes intertranslatable with time, work, money, part of "the economy"...a commodity that can be measured, and thence bought and sold'. (Gee, 1996: pp. 122-123)

The instrumentalised view of literacy that Gee presents is well acquainted with a narrow set of skills that enable the development of work-ready skills, something that the FAVE sector as a whole has, incorrectly, been characterised as being principally responsible for in recent years (Bathmaker, 2013).

But other definitions of literacy exist, and situate it as a cultural and social practice that plays a critical role in how an individual actualises themselves through both comprehension and language production in equal measure (Gee, 1996; Lea, 2004; Duckworth and Brzeski, 2015). Freire (1973) argues that literacy could act as a liberator of the individual through the attaining of a critical consciousness, through which one achieves an in-depth understanding of the world. These schools of thought invoke the deep held relationship between language production through writing, and how individuals make meaning of the world. Such links are valuable, and should not be lost. It is important to recognise that literacy within the FAVE sector has in recent decades erred on the side of functionality, through both policy and curriculum design, but that the subject itself comprises more than the sum of these parts. Significantly we can note that the adoption of GCSEs within FAVE contexts in 2015 led to the first formalised requirement for creative writing, and other broader curriculum elements, to be taught to 16-18 year old students in the sector, although the motivations for this shift (presented in Chapter One) were more aligned to the currency the qualification holds in employment and education markets than the curriculum content per se. What we now see is a sector that has positioned English predominantly as a reductive form of literacy through consecutive policy papers for the past twenty years, and that must now embrace the full breadth of English as a discipline, including creative writing. The implications of this are sizable as many English teachers based in the FAVE sector will have taught the subject amidst this legacy of changes, and this will have invariably shaped their experiences, knowledge and teaching practice of the subject, and will accordingly have imprinted themselves on teachers' assessment practices.

So what of the newly broadened English curriculum that now sits within the FAVE sector? The GCSE English Language qualification was first introduced in 1988 alongside the then newly created National Curriculum. The reasoning for the adoption of a National Curriculum can be traced back to the economic and political climate of the UK during mid-1970s. Public perception of education was falling in esteem as the belief that schools were not preparing students for the changing needs of industry and society became increasingly more pervasive. In response Prime Minister James Callaghan ventured the idea of a national 'core curriculum' and established consultations that ultimately led to the establishment of the National Curriculum as was introduced in 1988 and despite several revision and amendments, is still in place today. In light of the above, it is observable that the National Curriculum represented a seminal moment in the governance of how English was and still is taught in the United Kingdom. It 'provided a 'basic curriculum' to be taught in all maintained schools and [...] set out 'attainment targets' - the knowledge, skills and understanding which children would be expected to have by the end of each key stage' (Gillard, 2011).

On the international stage, countries typically structure their national curriculum around aims and values, subject content and skills, but do so in varying levels of detail. The UK National Curriculum, unlike some of its international counterparts, remains relatively prescriptive (House of Commons National Curriculum Report of Session 2008-09, volume I:9). The 2014 revision of the National Curriculum states that "English has a pre-eminent place in education and society" and "a high-quality education in English will teach pupils to write fluently so that they can communicate

their ideas and emotions to others" (2014:13). At Key Stage 4 pupils should be taught to "write accurately, fluently, effectively and at length for pleasure, [...] make notes, draft and write [...] and revise, edit and proof-read" (2014:19).

In all, the programme of study at Key Stage 4 comprises fourteen main standards that relate to student writing development, and over forty sub-standards that branch from these. On first glimpse it might be viewed as impressive that the entire discipline of writing has been neatly captured and defined in fourteen main and forty sub-standards that students need to meet, a feat all the more impressive considering the subjectivity and breadth of what writing constitutes. The reality is that such standards have been developed for what Sennett (2016) terms the 'massification of use', and have led to 'the objects themselves contain[ing] a mediocre level of functioning'. With the adoption of such standards, teacher agency, expertise and knowledge have been traded off in favour of prescribed and narrowed minimum expectations.

Such circumstances have important implications for the teaching and assessing of writing. Cremin and Myhill (2013) warn of the reductive nature of the Curriculum's content, and its side-lining of certain domains of writing in favour of others: 'writing has been conceptualised by some governments as little more than an unproblematic set of technical skills and has tended to become increasingly focused on writing outcomes, genre knowledge and skill mastery' (2013:1). Furthermore, D'Arcy (1999) suggests that 'since the advent of the National Curriculum, the way that teachers have been required to approach the teaching and the assessment of writing has become increasingly circumscribed within a narrowly mechanistic framework [...]

bound by a paradigm which focuses on writing largely as a matter of construction and correctness - at word level, sentence level and text level' (1999:3). With this in mind, I would argue that subscription to a hierarchy of prestige across different writing paradigms would invariably favour some students over others despite the fact that there are subjective merits in any form of student writing, and teachers must be mindful of this when assessing work both to judge performance and develop pedagogies of contingency.

## Alternative approaches to the assessment of writing

Although there was not a prescribed English curriculum of specific aims and outcomes before the National Curriculum was introduced, it would be erroneous to think that English played anything but a critical role in the education of individuals in the years predating it. Similarly, despite the absence of prescribed curriculum aims and outcomes it would be wrong to assume that there existed a general ignorance around what should be taught in English classrooms and how student writing should be assessed. In contrast, there exists a rich tradition of research, theory and practice on the assessment of writing that has evolved in recent times that has served to inform, influence and challenge ideas about what writing as a discipline is and how it should be understood and actualised in practical terms. Much of this work can be located aside from the prescribed summative assessment procedures mandated by the National Curriculum, and in some instances can be seen as reactions to counter its perceived shortcomings.

Before the adoption of the National Curriculum and the move away from 100 percent coursework weighting on GCSE English programmes, attempts to combine the assessment of both mechanistic and imaginative features of writing in practice had historically enjoyed some success. Britton's (1950) work into the marking of imaginative compositions sought to assess writing in a more holistic manner with criteria comprising pictorial quality and creativeness, although findings concluded 'we clearly did not agree on the qualities required of good imaginative composition' (1950:3). In a later trial on actual O-Level papers Britton (1964) employed multiple markers, in which three marked impressionistically and one for technical accuracy. He found that assessors valued writing that had 'real feeling' and 'real experience' (1964:23), and also identified a correlation in the impressions relating to textual involvement, organisation and mechanical accuracy assessors had across texts based on different subject matter (1964:27). Britton's work confirms the value of multiple marking of scripts by different assessors, but does not account for how they had reached their judgements.

In building on some of the recommendations of Britton's cited work, Wiliam (1994, 1996, 1998) advocates the use of 'construct referencing' where assessors award a grade and use a construct of what they think that grade looks like based on previous experience. In this 'the assessment system relies on the existence of a construct (of what it means to be competent in a particular domain) being shared by a community of interpreters' (Wiliam, 1998:6). The work of Britton and Wiliam here demonstrates the potential for other means of assessment outside of the standardised use of designated criteria. These examples serve to highlight teacher-led interventions in seeking to enhance pedagogy through the trialling of alternative writing assessment

practices centred on holistic appreciation of student texts and standardisation through collaborative marking.

Whilst the aforementioned studies advocate a collaborative approach amongst teachers when assessing writing, others advocate dialogic approaches that engender collaboration between teachers and their students. D'Arcy (1999) argues that the overabundance of summative marking criteria can lead to examples where 'criteria have been memorised as a check list' (1999:14), where there is little to no consideration made on how the writing impacted the reader and in rejection of this proposes an 'interpretative response' to text more akin to dialogic written feedback than conventional means. Within this mode of assessment, a reader that 'adopts a meaning-related paradigm would be prepared to take an aesthetic stance to the text, prepared to engage with it, imaginatively, empathetically, and visually' (1999:14). She forwards the idea that a teacher must engage imaginatively with their student's text, and not just mechanically, 'to assess achievement solely on the basis of the text's construction without taking its content into account seems at best inadequate and at worst absurd' (1999:15). This work resonates with the previously noted distinctions of assessment *for* learning as existing in the event of learning, and assessment *of* learning as existing after the event. For D'Arcy a dialogic approach to feedback is a way of sustaining interest and fostering development in the process of writing, rather than just the product.

Other research has explored how teachers' attitudes towards knowledge can impact the judgements that teachers make on student writing. Barnes and Shelmit (1974) surveyed teachers about the ways they use writing in class, and found that

responses could be grouped into two categories, *transmission* and *interpretation*. Responses in the transmission category suggested that some teachers saw writing as a means of students recording information provided by a teacher. In contrast responses in the interpretation category suggested that other teachers saw writing as a means of learning to think independently, to come to one's experiences or feelings, to construct one's meanings. Odell (1993) attributes this dichotomy as stemming from 'fundamentally differently attitudes towards knowledge', and suggests that:

> 'teachers who held an interpretative view were likely to see their job as allow[ing] students to explore ideas and deepen personal understandings of the world [...] by setting up a dialogue with students and encouraging them to use a piece of writing as a "springboard" for new individual or class projects' (1993:3-4).

We can perhaps reconcile the *transmission* attitude towards writing with Sfard's (1998) acquisition metaphor of learning, discussed in Chapter One, in which learning exists in tangible form and can be transported between teacher and student much like objects being passed from person to person. On the other hand, *interpretation* attitudes towards writing that Barnes and Shelmit present subscribe to what Sfard details as to other dominant metaphor to understanding learning: the participation metaphor. In this paradigm learning is viewed as 'evolving bonds between the individual and others' and 'makes salient the dialectical nature of the learning interaction: The whole and the parts affect and inform each other' (1998:6). This conceptualisation of learning aligns with the stances adopted by Vygotsky (1978) and Bruner (1986), who notes 'I have come increasingly to recognise that most

learning in most settings is a communal activity, a sharing of the culture' (1986: 127). These prevailing schools of thought cannot be easily resolved with ideas of individual measures of performance and highlight an apparent disparity between politicised state governed education and dominant theories centred on writing pedagogy.

These perspectives on writing, and on learning as a whole, call into question the validity of systems that seek to pin down and define what successful writing entails. What instead is suggested at is the establishment of an equitable, open and dialogic channel of communication to exist between teacher and student that comes about as a result of the production of work, much akin to the notion of an apprentice who through sustained application is beginning to understand the nature of their craft and a master who expertly guides them with expertise and encouragement. Moreover, also suggested here is the benefit that collaborating with fellow experts can yield for teachers when seeking to reach judgements on student writing. Both Britton and Wiliam advocate approaches to judgement that help teachers counteract what Marshall (2011) argues is their 'apparent distrust of 'analytical' forms of assessment [that] arises from the nature of the discipline" (2011:29).

At this juncture we are faced with a quandary. At one end of the continuum we can appreciate the discernible benefits of forming interpretive judgements of student writing that are not alone defined by curriculum standards, that are formed as a result of careful collaboration and consideration with other teachers and that also facilitate a dialogic mode of feedback. Moreover, it is suggested that teachers must be cognizant of the standards that exist beyond the codified standards of the mark scheme when assessing student work as they seek to form an effective judgement.

These standards beyond the codified might be termed tacit knowledge, a concept that will be discussed in greater detail shortly. At the other end we can note the prominence of institutional accountabilities, punctuated by management mantras such as all work must be marked that take up much of teachers' time and effort day to day and undeniably infringe on the adoption of such practices.

This thesis seeks to tentatively challenge the polarity of this continuum by exploring the impact and resulting effects of trialling a different approach to the evaluation of milestone assessments completed by a modest sample of 16-18 students. The approach draws from and builds upon from the research findings presented above. However, in doing so it does not mimic their specific approach or methodology. What translates into the adopted methodology is an appreciation of how standards can inhibit teachers when attempting to form judgements on the quality of student performance in creative writing tasks, and the benefits that can arise from cross-teacher collaboration when assessing. Student performance is judged, but not quantified through measurement, through the use of an Adaptive Comparative Judgement (ACJ) approach to assessment. Further discussions of the methodological approach adopted in enquiry feature in Chapter Three.

# Adaptive Comparative Judgement

## Comparative vs. absolute judgements

In view of the above discussions, the challenge of aligning student work to assessment standards still persists, even if we accept that they are solely a heuristic to be used in tandem with the teacher's experience, skills and tacit knowledge to help form judgements. In such circumstances, in which it is not always clear to a teacher what they should do and where they might not have all of the information they need to make an informed judgement decision, issues of validity, accuracy and reliability arise. These issues derive partly from the fact that the teacher is required to assess each item of student work in isolation from one another rather than with an appreciation of how other responses compare with it. But alternative approaches to assessment do exist, albeit with relatively low exposure in wider educational circles. This research focuses on one of these alternatives in some depth.

Adaptive Comparative Judgement (ACJ) is an assessment methodology that offers an alternative approach to the conventional approach to individualised criterion referenced judgement. Derived from the research of Louis Thurstone featured in *Law of Comparative Judgement* (1927), in which he argued that while humans have great difficulty making quality judgements with validity and reliability we are much more adept at making comparative judgements - judgements of quality between two items. ACJ differs from conventional modes of assessment, and what is frequently termed absolute judgement, in which scripts are read and assessed in isolation from its counterpart scripts against predetermined criteria. In advocating comparative judgement approaches, Laming (2011) argues that 'There is no absolute judgement.

All judgements are comparisons of one thing with another'. Pollitt (2012a) extends this idea in stating:

> 'When a judge is asked to make an absolute judgment about the perceived quality of an object, previous experience, level of knowledge, self-efficacy and the opinions of others all influence that judgement. In summative assessment, examiners are (and should be) greatly influenced by the mark scheme to an extent that overcomes bias as far as possible. So, the absolute judgement of what mark to award is relative to the mark scheme plus any error and bias in its interpretation.' (2012a:2)

The resolution to such challenges, it is proposed, is in the shifting in focus of what the judgement is made against. Rather than locating the quality of individual objects against a scale of quality, ACJ is only interested in the difference in quality between the two objects. In such an approach the only requirement for the judge is to be able to perceive the difference in quality based on their own personal standard or external criteria.

As already discussed, there exists much debate on the natures of and relationship held between explicit vs. tacit knowledge, and how different conceptions of what comprises a 'quality' item of work exist in a discipline as open-ended and subjective as creative writing. One significant feature of ACJ that attempts to address this matter is the flexibility it permits to judges regarding through what lens quality of items is judged through. ACJ is not solely reliant on a mark scheme to guide the judgement decision and offers teachers far more agency in considering what a 'quality' item is. This is possible as the comparative nature of the approach provides

the judge with a frame of reference on which to base their judgement regardless of the focus. This differs from absolute referencing, in which the judgement decision is entirely contingent on their interpretation of assessment standards. Bartholemew (2017) observes with an ACJ approach that a judge's decision can be based on 'viewing two items of and choosing the better of the two based on their own expertise and predetermined criteria or rubric' (2017:2). This suggests that decisions can be arrived at through a combination of a judge's conception of quality and predetermined criteria. Such combinations may be well placed to act either as an explicitly standardised definition of quality, or as an interpretive heuristic.

In an ACJ mode of assessment, each script is seen several times in different pairings to develop a ranked order of performance across a series of student scripts (Pollitt, 2004:6-7). Over time student work gains a "win-loss" record; each time an item is chosen over another the piece of student work it counts as a "win," while a "loss" stems from not being chosen when paired with another item (Pollitt, 2004, 2012a; 2012b). After a number of different pairs have been judged in various combinations a ranked arrangement of scripts begins to form. Recent advances in technology have led to the creation of ACJ software applications and online platforms that can help to facilitate the process. With software teachers can view two scripts on a computer screen and choose the better of the two, making the process much more efficient than with paper-based approaches. This software also helps ensure that ACJ approaches to assessment are truly 'adaptive' and respond to the decisions that are being arrived at. In recent years, ACJ has been piloted, tested, and refined over time and the algorithm which facilitates the judgments has been improved (Pollitt, 2004; 2012b). As resultant pairings are increasingly more refined,

and rather than random pairings being presented, items of work that hold similar 'win-loss' records are compared with one another and the overall rank in improved in terms of validity and reliability (Pollitt, 2004).

## ACJ as assessment of learning

Several studies have used ACJ approaches in teachers' assessments of students' work in open-ended tasks and performances. Heldsinger and Humphry (2010) focus on a study involving twenty staff from a school in Australia who judged thirty narrative texts from students aged six to twelve years old with the intention of ascertaining the viability of alternative approaches to assessment away from large scale testing programmes. The findings report a high reliability of the rank order (0.982, derived from the Rasch model of analysing categorical data). Comments from staff note that the process 'force[d] consideration of the qualitative characteristics that distinguish one performance from another' (2010:16). Also noted by these researchers is that some teachers involved in the study perceive the ordering the scripts from lowest to highest in a scale provided valuable information for future teaching programs by characterising the zone of proximal development of students. The reference here to the zone of proximal development (Vygotsky, 1978) has much in common with concepts of assessment for learning. Whilst this was a likely unexpected finding from the study and was not pursued further by the researchers, it remains a point of interest from this study in view of our exploration of how to reconcile formative and summative modes of assessment when judging quality of work.

Kimbell et al. (2009) led a study of 28 teachers of design and technology, geography and science in using ACJ to assess performance in GCSE work, and chose to adopt this approach 'because we are essentially concerned with assessing performance-based capability, an overview holistic judgement of the performance seemed more appropriate.' (2009:13). The findings determine a high reliability rating of 0.95 using Rasch analysis. Interviews with teachers noted that 'it still took me some time to overcome ingrained, detailed examination of the folios and being prepared to adopt the holistic judging system. This became much easier after having completed a number of comparisons' (2009:154). The study concluded that the use of a full day of training for the participating teachers, which centred on forming comparative judgements and in using the online software that would facilitate this, was sufficient in equipping the participants with the ability to make paired comparisons.

Other research has sought to determine the viability of using ACJ for summative assessment purposes. Whitehouse and Pollitt (2012) recruited 23 teachers of AS level GCE Geography, of whom sixteen had experience examining for an awarding body and the remaining seven had no experience of, to comparatively judge a sample size of 564 essays. A total of 3500 paired comparisons took place. The study chose not to use assessment standards, instead providing two *importance statements* to teachers when judging. The first comprised the aims of the AS level specification in geography, which offered a link between the rigour of the GCE specification and making holistic judgements without a mark scheme. The second comprised the following statement: '*Based on these statements, which of the essays*

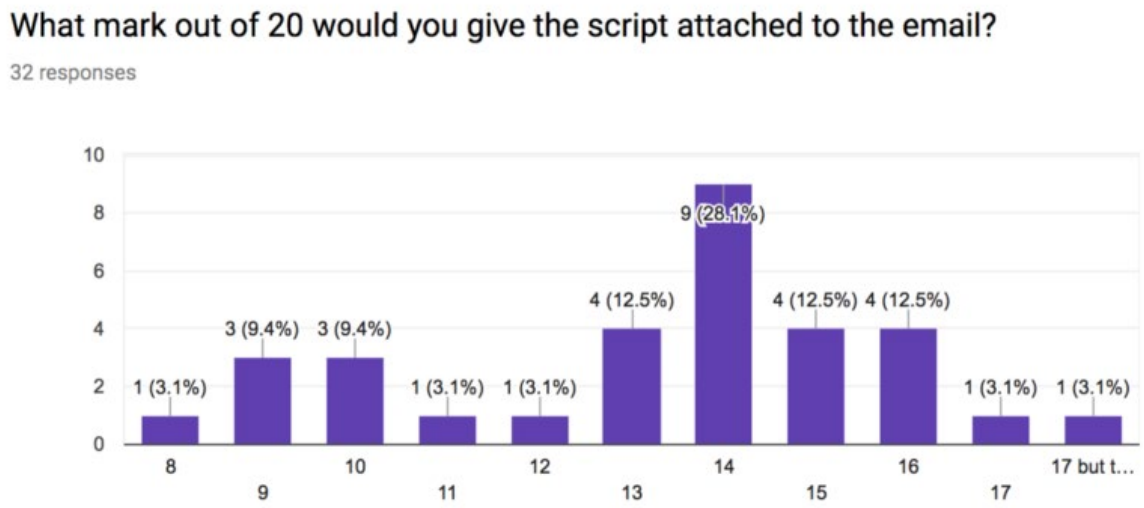
*shows more evidence of a higher level of development of what is deemed important in Geography?'* (2012:6).

Findings determine that the process produced a high level of reliability in the judgements that were formed (0.97 using Rasch analysis). Despite this it was found that some judges experienced difficulties in not using a mark scheme, and others did not use the importance statements in favour of relying on their own professional instinct. It is concluded that 'it is insufficient for an awarding body to offer no guidance on how to assess its high stakes exams, thus further work is required to find out what sort of guidance is most effective at the point of judgement' (2012:15). This is a worthy point, in that awarding bodies are of necessity required to define standards and provide guidance for qualifications they publish for summative purposes. Despite this, the findings here suggest that teachers can successfully draw on professional knowledge and experience to help inform them of the relative quality of work in lieu of using highly prescriptive assessment criteria when they adopt an ACJ approach to assessment, and still form reliable judgements as a result.

The last few years has seen an increase in interest and exposure to ACJ approaches to assessment in the wider educational landscape, which can in part be attributed to Pollitt's (2004, 2012a, 2015) sustained interest in using the approach in varying contexts. This has resulted in a renewed interest in investigating alternatives to prevalent absolute judgement assessment models, in some circles. One initiative that has arisen as a result of this is NoMoreMarking, a national organisation that leads projects on the use of ACJ on English study programmes in primary and secondary schools. They currently lead projects on the assessment of writing in

years 1 to 6, on the assessment of reading and writing in years 7 to 9 and on assessing GCSE English in years 10 to 11, and institutions can subscribe to these to gain access to the training, standardisation and moderation offered by NoMoreMarking. The organisation's website, NoMoreMarking.com, provides free access to ACJ algorithm software through which teachers can upload work and judge its quality in adaptively refined pairs.

Small-scale research conducted by NoMoreMarking (2017) surveyed 32 of their affiliate school coordinators, comprising heads of English based in secondary schools, to provide a mark to a student's GCSE English reading answer script using the mark scheme. The original mark was removed, and the markers were not aware of the purpose of the exercise. The results are seen below:




*(Figure 2.1) - NoMoreMarking absolute judgement responses from GCSE English reading assessment*

The significant disparity in marks awarded seems to further reiterate the discrepancies that can arise as a result of absolute judgement against mark

schemes. The results do indicate some commonality in the judgement reached across these teachers, notably on the awarding of fourteen marks. This trend might point to a common interpretation of the mark scheme's standards by those teachers and a misinterpretation by the others, although we can only hypothesise.

The previous study serves as a preliminary enquiry to a larger scale study that NoMoreMarking undertook in 2018. This enquiry seeks to explore if teachers use comparative judgement to judge rather than mark GCSE English writing mock papers could this reduce workload, and to find if teachers would agree with one another in their judgement decisions to a high level of reliability. Overall, 37 schools and 396 GCSE English teachers participated to judge 5530 student essays. Teachers were given no specific guidance on how to judge and were instead presented with the question: 'The better writing?' Anchor scripts were added to the student scripts, unbeknown to the judges, that were moderated samples of work at specific grades and levels. This allowed NoMoreMarking to determine at what position in the rankings specific essays were working at, at to facilitate the comparison of results.

Regarding workload, the findings determine that the median judgement time for the writing essay was 23 seconds per pair, and that a typical scenario 'appears to be a teacher spending just over half an hour to judge one set of essays' (NoMoreMarking, 2018). In reliability terms the findings suggest that after a script had been judged alongside a counterpart script in at least fifteen different pairings the rankings that followed were outperformed the marking metrics reported by Ofqual, of +/- 5 for a 40

mark essay. Across the sample of 5,530 scripts the variation in marks in the writing scripts was just under 2.8 marks.

The findings here are hugely encouraging, and do seem to suggest that ACJ offers a viable and practical means with which small and large-scale assessment can be undertaken. However, it might be argued that there are some shortcomings in the approach that NoMoreMarking has taken. The intention of their projects is to identify viable alternatives to assessing work through absolute judgements, but their work is very much aligned to fulfilling the demands of assessment *of* learning. It might be argued that they are exploring more practical and accurate alternatives of assessment practices that can meet the accountability demands imposed on teachers by institutions. Certainly, since the inception of NoMoreMarking in 2015 this initiative has acquired significant backing from schools nationwide that seem happy to at the very least trial this mode of assessment. Despite this, the matter of assessing students solely for the purposes of ranking their performance in designated windows throughout the academic year to then report on progress still does very little to contribute to an assessment *for* learning. Whether a student is working at a 'grade 6' or is 'ranked 12 of 32 in their cohort', the result is the same for the student who likely has no conception of what this actually means in terms of learning, progress or their next steps. Research that works with teachers in establishing how ACJ can lead to assessment for learning remains an under researched area.

The results presented by NoMoreMarking do present another potential matter of concern that should be noted. The findings present ACJ to be an intuitively appealing

mode of assessment in that student work can be judged at a much faster rate than through the conventional absolute judgement marking system. The declaration that a whole class of essays can be judged within a good degree of reliability by a teacher in only thirty minutes is undoubtedly an eye-catching claim sure to be of interest from stakeholders across education. But focusing too much on possible time saving ACJ offers and not on the actual process of judgement can risk diminishing the professional practice of assessment to a mere quantifiable output. Institutions must consider their motivations for the adoption of different practices; in a scenario in which an institution adopts ACJ but finds it to take longer for teachers to judge in this manner that with absolute criterion judgements, one wonders if it would it be retained, regardless of how reliable the process was. Sennett (2008) likens standardising reforms adopted by the National Health Service in the mid-2000s to 'Fordism', which 'takes the division of labour to an extreme: each worker does one task, measured as precisely as possible by time-and-motion studies; output is measured in terms of targets that are, again, entirely quantitative' (2008:47), and similar risks present themselves for ACJ here. ACJ is currently enjoying relatively high exposure in primary and secondary school settings, and it remains detached from institutionalised accountability mechanisms. Whether an increasing influence in the ways institutions report student progress would lead to ACJ losing its educational value through the neglect of genuine judgement practice is not yet known, but is something that teachers must be aware of.

## ACJ as assessment for learning

The vast majority of research into ACJ approaches to assessment has focused on what can be learnt from using the approach with teachers, but there is some research focused on using ACJ with students. Hardy et al. (2015) employ ACJ to facilitate peer assessment activities that were implemented in undergraduate courses in physics (231 students) and pre-clinical veterinary medicine (154 students). In both scenarios the ACJ assignments were based on 'long answer' questions from previous exam papers and were chosen to enable students to compare their performance with that of their peers, something they would not normally have the option to do. In addition to forming a comparative judgement, students were asked to provide a short feedback comment for the author of each submission that they encountered. Student submissions were marked by academic teaching staff to allow for comparisons to be drawn between student rankings and the marks teachers had assigned (Hardy, 2015). For physics, it was found that there was no correlation between the quality of assignments based on student ACJ rankings and numerical marks awarded by staff, but in contrast, significant correlation was found in veterinary medicine. This was attributed to physics students not having access to explicit assessment criteria and lacking confidence in their own subject knowledge and judgements. Conclusions noted that 'this demonstrates the importance of expert guidance to help students develop their assessment expertise...and that opportunities for practice coupled with timely feedback are also needed' (2015:18-19).

The student-centric approach to ACJ adopted by Hardy et al. is one that is in need of further investigation to fully realise its potential effects with students across different

contexts and disciplines. This study comprised higher education students undertaking ACJ. From this we can take that they were perhaps more mature, autonomous and intrinsically motivated in their approach to the exercise than students in other sectors of education might be. Nonetheless, there are features that are unique to this enquiry when compared to the other ACJ research that has been presented. With reference to wider literature on peer-assessment, the findings from Hardy et al. align with Rust, Price and O'Donovan's (2003) findings that show that an intervention aimed at improving students' 'assessment literacy' through explicit assessment criteria and tacit knowledge resulted in improved performance. Moreover, Falchikov and Goldfinch (2000) show there was a significant correlation between peer and staff marking, with strongest agreement when the assessment involved global judgement using well understood criteria.

The mention of global judgement and tacit knowledge chimes with Sadler's (1989) conception of *guild knowledge*, that comprises 'the ability to make sound qualitative judgments', that are forged through 'a history of previous qualitative judgments and where teachers exchange student work among themselves or collaborate in making assessments' (1989:126). Sadler's argument is that a teacher's guild knowledge should consist less of knowing how to evaluate student work and more 'knowing ways to download evaluative knowledge to students' (1989:141). If we are to accept that the study of English comprises largely subjective interpretations of knowledge that are difficult to define in standards and mark schemes, this carries significant implications for what pedagogies teachers can viably utilise. Sadler's assertion is that, much in the same way that teachers must be inducted into communities of learning through which they can draw on pre-established practices, traditions and

knowing to build their own tacit knowledge of a discipline, so must students. Having students undertake peer assessment, through which they must contend with the negotiation of assessment standards, offers a viable way of doing this.

In research into assessment practices in English teaching, Marshall (2011) observe that ‘teachers were able to share with their pupils some sense of ‘guild knowledge’ in the process of writing assignments and they did it predominantly through peer assessment’. Conclusions that were drawn following interviews with teachers noted that through sustained use of peer assessment activities ‘the class gained ‘good knowledge’ of ‘quality’ that cannot be expressed in a tick box. They have moved from counting a variety of sentences to recognizing that ‘quality’ is something more [...] Writing in this sense has become more of an abstract, more of a concept that can be seen in many ways’. (2011:101). Despite tentative findings of Hardy et al. (2015) that advocate the use of ACJ as a peer assessment method, the possible benefits in using such an approach with English students within FAVE contexts is not yet apparent. From Marshall's work we can take that peer assessment can provide students with conceptions of what ‘good’ quality work looks like, beyond standards and tick boxes. Whether ACJ as a peer assessment method might engender a broader holistic appreciation of varying script qualities is worthy of further enquiry, based on what has been presented. This is the primary focus of this thesis.

# Chapter 3 – Methodology

This Chapter presents the methodological approach underpinning this enquiry. It begins by presenting the research questions, including a detailed examination of the possibilities and unexpected findings that might arise through the investigation of this research problem. This introductory section also includes a rationale explaining how the focus of this research came to light in practice.

The Chapter then addresses wider methodological considerations, notably the researcher's position, and a justification for the research as a pragmatic necessity that has arisen from grounded practice-centred experiences. Discussions centre upon ontological and epistemological issues considered in relation to the research problem, and ethical considerations following from the above are made. The Chapter concludes with a presentation of the data collection and analysis methods the research employs in this study.

## Research questions

This enquiry has one main research question it is seeking to address:

***RQ1:*** *What are the benefits and challenges of using an adaptive comparative judgement approach when assessing GCSE English creative writing scripts in a Further Education institution, as perceived by me as a practitioner researcher?*

From this, there are three subsequent research questions that follow:

***Sub-RQ1:*** *What new knowledge can be acquired by the teachers involved in the enquiry as a result of undertaking adaptive comparative judgement and what function does this serve them as teachers of GCSE English in an FE context?*

***Sub-RQ2:*** *How can adaptive comparative judgement be used across a team of teachers to standardise assessment practices?*

***Sub-RQ3:*** *What can learners' adaptive comparative judgement decisions tell us about their understanding of creative writing as a field of study in the discipline of English Language, and what are the subsequent pedagogical implications that follow from this?*

Trochim (2000) suggests that there are three types of research questions:

- Descriptive questions - that aim to explore or to describe what is currently taking place.
- Relational questions - which seek to determine associations between linked objects.
- Causal questions - which determine whether one or more variables lead to specific outcomes

(2000:25)

Each of the questions that underpin this enquiry are what Trochim terms descriptive in nature, in that they are seeking to explore issues from a relativist position and provide insight following from this. This categorisation is helpful in clarifying the macro-level focus and intention of each of these questions, but to suitably frame this research a more substantial analysis and justification of each question is required. This following section of this Chapter comprises an explication of each of these questions, including an examination of what I am trying to find out and why each of these questions represent areas of interest in this enquiry. The section concludes with an overarching discussion that will sketch the commonalities in theme and scope that these questions share.

## Research question 1 (RQ1):

***RQ1:*** *What are the benefits and challenges of using adaptive comparative judgement approaches when assessing GCSE English creative writing scripts in a Further Education institution?*

The construction of this question has been designed to invite a critical examination of the use of comparative judgement approaches in the practice of assessment. Adaptive comparative judgement is an approach to assessment that has been demonstrated in research to provide benefits for educators across ranging contexts; these include an increased reliability in assessment decisions (Heldsinger and Humphry, 2010; Kimbell et al. 2009;), a significant reduction in the time it takes to assess each script (NoMoreMarking, 2018), and the providing of opportunities for teachers to draw on tacit knowledge beyond those of codified standards in making

their assessment decisions (Whitehouse and Pollitt, 2012). With this said, there is a paucity of research into the possible impact of using adaptive comparative judgement in FAVE settings. The methodological approaches adopted by the research cited above serve as framing tools for the possible areas of focus within this enquiry. Indeed, it was this research, and other examples, that first drew my attention to ACJ as a possible viable alternative to traditional assessment and that led me to explore this in my own context. However, this enquiry must be responsive to the likelihood that the trialling of this approach will experience challenges as well as benefits. From an ontological position, to assume that trialling this approach in an unfamiliar context will yield similar effects to those that have been reported in other research contexts would be erroneous, in that this would be assuming a positivist absolutism about the impact such an approach can have in any given setting.

In terms of scope, this question has several possible lines of enquiry. These largely follow from the research cited above that have focused on the application of ACJ assessment approaches in different contexts. The first of these will be to determine the average time spent assessing student creative writing scripts using an adaptive comparative judgement approach. Following from is the second line of enquiry, which will determine the reliability and accuracy of the assessment judgements that are being arrived at through use of an ACJ approach. This is crucial in helping to determine that the findings from the first line of enquiry are valid, in that the central tenet of any assessment judgement should first be reliability and accuracy before considerations of time invested per assessment judgement are considered.

The third will be to explore the impact of giving practitioners opportunities to assess student creative writing beyond codified standards when making their assessment decisions. ACJ, as noted by Whitehouse and Pollitt (2012), is an assessment methodology that is particularly well-suited to giving teachers an impetus to draw on tacit knowledge when judging student proficiency in an assessment context. The benefits and challenges pertaining to this line of enquiry will in-part be determined through the cross-referencing of findings against the previous two lines of enquiry, in that assessing through an ACJ approach beyond codified assessment standards might have a discernible impact on the time taken for, and reliability of judgements. But in order to fully explicate this it is also important to ascertain the thoughts and opinions of the teachers that are using this approach. These insights serve to elucidate the individual experiences of each teacher, be these positive or negative.

## Sub-research question 1 (Sub-RQ1):

***Sub-RQ1:*** *What new knowledge can be acquired by teachers as a result of undertaking adaptive comparative judgement and what function does this serve teachers of GCSE English in an FE context?*

While the impact of applying ACJ approaches to assessment has been an area of some research interest in the past fifteen years or so, there remains within this domain a distinct lack of attention paid to the role that ACJ can offer teachers as a means of acquiring new knowledge. In this question I am positioning new knowledge as a knowledge of both assessment as a practice and of subject content. As explored

in Chapter Two, the prevailing tendency to frame assessment as a form of measurement in educational circles is not easily reconciled with Laming's (2011) assertion that 'all judgements are comparisons of one thing with another'. Through this question, this enquiry aims to explore the impact of using ACJ approaches with teachers to determine how the explicit framing of assessment as a practice in which a judgement is made through comparison, rather than an isolated measurement, impacts on teachers' knowledge of both the process of assessment and the subject content with which they are engaging.

## Sub-research question 2 (Sub-RQ2):

***Sub-RQ2:*** *How can adaptive comparative judgement be used across a team of teachers to standardise assessment practices?*

This question speaks to a practical dimension in respect of the use of ACJ approaches to assessment. As noted in Chapter One, the problem of wildly varying assessment judgements that provided the initial impetus and contextual backdrop for this enquiry can be characterised as a need to better standardise the assessment decisions that teachers are arriving at. Standardisation in this respect is positioned in different forms: to external quality markers, and to other teachers' judgements as socially-situated tacit understandings of what makes a good piece of creative writing. It is intended that ACJ will provide a method through which both of these types of standardisation can be achieved.

## Sub-research question 3 (Sub-RQ3):

***Sub-RQ3:*** *What can learners' adaptive comparative judgement decisions tell us about their understanding of creative writing as a field of study in the discipline of English Language, and what are the subsequent pedagogical implications that follow from this?*

This question seeks to address one of the predominant shortcomings of research that explores the use of ACJ; that ACJ is an approach to assessment in which teachers act as the judge. An under-researched but pertinent point of interest that this enquiry seeks to explore through this question is how ACJ can be used with students as a mode of peer assessment. Much literature around what constitutes effective formative assessment maintains the importance of learner ownership of the learning process, with assessment comprising an integral part of this (Sadler, 1989; Marshall, 2011; ETF, 2017). This line of enquiry seeks to explore how engaging learners with ACJ as peer assessors can inform us of their understanding of creative writing, and whether it can act as a pedagogical approach through which 'guild knowledge' (Sadler, 1989; Wiliam, 1998), as discussed in Chapter Two, can be developed.

# An introduction to practitioner-led research

## Practitioner-led research

This is a study in practice-focused educational research. It does not set out to test a hypothesis, or to justify the use of a specific set of research methods. Rather, as Armstrong and Moore (2004) state, it aims to "carry out the evaluation of a particular intervention which has an identifiable focus and purpose, but which does not predetermine outcomes, or discard those that are unexpected" (2004:2). McNiff and Lomax (2004) observe that one of the prevailing reasons for educators engaging in practice-focused research conducted by front-line practitioners is to "investigate what is happening in their particular situation and try to improve it. They not only observe and describe what is happening; they also take action." (2004:14). Authenticity of the situated context of the enquiry is crucial as this can provide insight that is actualised in respect of its own context, and subsequent issues the practitioner might encounter. The resulting action that is taken can also benefit from an appreciation of the context for this reason. With this in mind, it is important to state that while this enquiry takes action that seeks to address a perceived shortcoming in existing practices in relation to GCSE English assessment in my institution, the primary focus throughout is upon an exploration of the benefits and shortcomings of a new approach. In short, the aim here is not to 'prove' or 'disprove' the success of a hypothesis but to explore and present an authentic account of experiences of action taken in context.

In the conduct of practice-focused research, it is important that the researcher acknowledges their own positionality in relation to the context and focus of the investigation. This is particularly important in an enquiry such as this, where I am the

researcher and a colleague and/or a teacher of the participants in the research. The quality of relationships forged between the researcher and the research participants before the commencement of this enquiry, impact on the interactions, and the data gathered from interactions, that occur between parties. Moreover, it is also possible that my own position and my own experiences as an insider in the research, might lead to a misinterpreting or misrepresentation of research data due to the high degree of familiarity that I have developed within the context of this research. In order to reduce this, measures have been taken in the research design to triangulate data sources in order to increase the authenticity and trustworthiness of the findings of the research. In order to achieve this, deliberate choices in the design of the research have been made. These include, the processes employed in the recruitment and selection of research participants, methods of data collection and methods of data analysis. These are discussed later in this chapter in relation to justifying the choices that I have made as the researcher. It is also important to note that being an 'insider' in the research also has advantages as well as disadvantages. For example, as an 'insider' I may be able to notice the symbolic significance of phenomena that an 'outsider' researcher might overlook.

In reference to practice-focused research, Coffield et al. (2004) invite one preliminary consideration: "Before making any change in practice, professionals are duty bound to consider two possibilities: first, that the proposed change may make matters worse; and second, that some alternative change may be more beneficial than their preferred option." (2004: 135). This enquiry aims to illuminate some of the issues raised in Chapters One and Two, and to offer insights into the research questions posed in Section Three of this Chapter, whilst simultaneously reporting research

findings in proximity to the context in which they were found. When reporting findings, the intention to offer what Bassey (2003) terms 'fuzzy generalisations [...] which suggest that, for example, *it is possible* or *it may be in some cases* or *it is unlikely*' (2003: XI). Underpinning this is what Bassey describes as a best estimate of trustworthiness, "a professional judgement based on the experience and reading of the researcher. [...] Making a best estimate of trustworthiness demands that the researcher thinks about the empirical findings of a research project in terms of who may use it - and how useful it may be to them" (2003: 1). While the intention of this research is to seek and unearth new knowledge, it is not the aim of this study to provide conclusive statements on how assessment practice, and other associated practice, can be improved through replication of the methodological approach that this enquiry adopts. Rather, in setting out the findings and recommendations that feature in Chapters Four, Five and Six I encourage readers to identify parallels and contrasts between their experiences of assessment practice and what is presented here. This point about locating oneself in a time and space relative to the context of this research to gain an understanding of its nature, including where it has been conducted, who is involved and when is it taking place, lead to considerations of research paradigms.

# Research paradigms

## Locating a research paradigm

The intention of this enquiry is to gather a body of empirical evidence with which new knowledge might be uncovered, with the intention of reaching 'fuzzy generalisations' about how assessment practices in the FAVE sector might be done differently. But, on the matter of forming an empirical base of evidence there is a need to consider and define the approach to understanding what these sources of information will be able to tell us, about assessment practices, about teachers and students experiences and the ways they think and learn, and the world in general. Kuhn (1970) describes a person's conception of the world, its nature and their position in it, as well as a multitude of potential relationships with that world and its constituent parts as a paradigm; 'as a world view or perspective – being shared by groups of researchers who adopt the whole paradigm as the one true way and defend it in opposition to any other set of views' (1970; cited in Coe et al., 2017:5). Conflicting ideas about how the world can be seen and understood have enjoyed varying levels of prominence in the field of educational research in the past few decades (Coe et al., 2017; Waring, 2017).

Waring (2017) sets two of the predominant paradigms that are hallmarks of educational research: positivism and interpretivist at each end of a continuum where positivism is located on the left and interpretivist on the right. For Grix (2002, 2010 cited in Waring, 2017:15-17) educational research comprises four 'building blocks', which he identifies as ontology, epistemology, methodology and methods (see Figure 1, below). He argues that in combination with one another these building blocks

inform the particular paradigm to which the research, and researcher, aligns themselves. The subsequent discussions in this section of the Chapter initially draws a contrast between positivist and interpretivist paradigms, before considering the stance that this enquiry adopts in relation to the first two of Grix's four building blocks of educational research. Discussions later in the Chapter present the methodological approach, and methods employed.

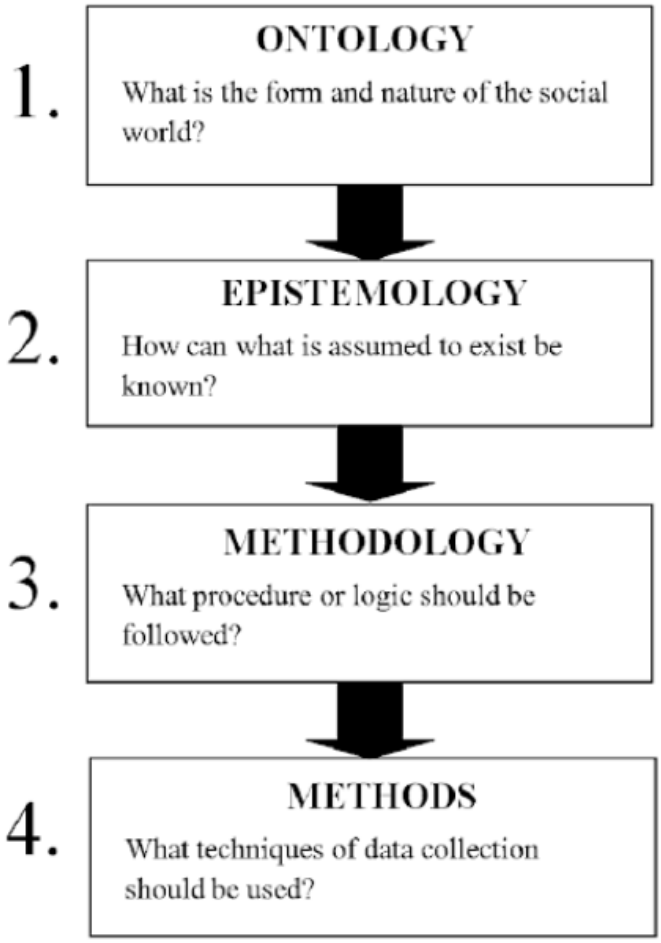



*Figure 1: the relationship between ontology, epistemology, methodology and methods (Waring, 2017; adapted from Grix, 2002 & Grix, 2010)*

When the term paradigm was first used by Kuhn in the late 1960s and 1970s there existed a dominant view in educational research that the scientific perspective, which favoured hypothesis and statistical-testing approaches, was considered to be the

most ‘robust’ the most valued (Coe et al., 2017). With reference to the FAVE sector, this view has in the years since shifted somewhat, and whilst scientific approaches to educational research still enjoy some large-scale national support, notably by organisations such as the Education Endowment Foundation (EEF) who in collaboration with the Sutton Trust receive substantial financial backing from the DfE (EEF:2019) to conduct hypothesis-centric educational research, there is however now a broader tradition in the Further Education landscape that is appreciative of other research paradigms (NetworkingtheNetworks:2019).

## Ontological Considerations

Research paradigms which regard themselves as following scientific traditions tend adopt a positivist ontology, in which objective truth about the world is not only considered to exist but is capable of being unquestioningly established through rational inquiry. For positivists, Carr (1995) observes, ‘educational inquiries are simply scientific inquiries designed to improve the rationality of education by purging it of any dependency on irrational dogma or subjective belief’ (1995:112). In positivist traditions, researchers explore the world impartially, discovering absolute knowledge about an objective reality (Scotland, 2012:10). From positive perspectives, history and context are often detached from knowledge, which is viewed as being objective and absolute. Following from this, language is seen as a constant operating in a representative role and words owe their meanings to the objects which they name or designate (Frowe, 2001:176).

Positivist paradigms seek to provide clarity and define the limits of certainty in the research findings they report. However, some thinkers argue that educational research that subscribes to a positivist paradigm cannot reasonably be considered to comprehensively and consistently offer 'truth' about the world (Scriven, 1970; Berliner, 2002), and that absolutist views of the world as advocated by positivists can pose some difficulties in educational research contexts. The close link between cause and effect that is emphasised and tested in positivist paradigms can be susceptible to not appreciating all variables that are present and that might be affecting the outcome in some way including perhaps the most important variable of the experiences and the fallibility of the human being conducting the research. Accordingly, the problem of causation or correlation becomes a concern, in that a researched intervention might have impact on resulting actions or behaviours. This challenge is amplified when applying the positivist paradigm in educational research, as the subjects of the research are highly variable as are the contexts in which they interact.

On this, Wiliam (2019) draws a contrast between research in traditional scientific domains, such as Physics, and educational ones, observing that:

> '…the problems that teachers need to solve are just much harder. Physics works because protons and electrons don't have good days and bad days; they behave consistently, and predictably. As soon as humans are part of the picture, things get a lot more complicated.' (2019:TES online)

The complications that arise from conducting research with humans that Wiliam points to appear to point to the value of interpretive paradigms to be a consideration

alongside positivist paradigms in educational research. Stables [1996] supports this view and suggests that:

> "One of the advantages of developing educational research beyond its original empirical positivist tradition has been a broadening of its subject matter; another has been its increased potential to call forth different kinds of reading." (1996:9).

The 'broadening of [...] subject matter' within interpretivist traditions discussed here helps to bridge the complications that Wiliam cites, in that interpretivist approaches permit and even encourage unexpected, unforeseen and difficult to understand insights that would remain hidden in a positivist paradigm. A research paradigm that subscribes to a more interpretive approach gives the researcher liberty to explore ideas of ontology and epistemology more freely, in that it acknowledges the subjective nature of reality in which differing and contrasting interpretations exist. A broad and uninhibited understanding of these notions is integral if we are to value and respect each individual's version of reality as being able to provide an illuminating insight into the matter at hand.

From an ontological perspective, interpretivist paradigms support the idea of locally constructed versions of reality. As Waring observes, 'we cannot see the world outside of our place in it' (2017:18). Knowledge and reality are constructed through interaction between humans and their world, and are developed and transmitted in a social context (Crotty, 1992:42). Therefore, the social world can only be understood from the standpoint of individuals who are participating in it (Cohen et al., 2007:19). In research, these aspects yield both positives and negatives; interpretive paradigms

are sensitive to individual meanings that can help form generalisations that draw from the collective voice, but without reference to an external anchor, as might be found in a positivist tradition, questions of legitimacy and trustworthiness in any consensus that is reached must be addressed (Scotland, 2012:12).

Waring (2017) offers researchers the following opening question that needs to be asked when considering matters of ontology in educational research:

> *'What is the nature or form of the social world?*
> (Waring, 2017:16)

From the above discussions, and in consideration of the content of Chapters One and Two, an interpretive ontology aligns most closely with the aims of this research. The intention of this research is to explore experience of action taken in context, rather than examine/prove/disprove a hypothesis. As such, it follows that adopting a research stance in which individual versions of reality can be documented and evaluated in respect of the context in which they are situated, and with a recognition that the world does not exist in absolutist terms is justifiable in this study.

## Considerations into epistemology

Grix's (2002, 2010) second building block of educational research is epistemology, concerned with the nature and forms of knowledge. Epistemological assumptions account for how knowledge can be created, acquired and communicated, in other words what it means to know (Scotland, 2012:9). When considering matters of

epistemology, it is natural to follow from the position of ontology that has already been determined. Indeed, Waring (2012) observes that in an interpretivist ontological paradigm:

> '...the investigator and the object of the investigation are assumed to be interactively linked so that the 'findings' are literally created as the investigation proceeds. Therefore, conventional distinction between ontology and epistemology dissolves' (Waring, 2012: 18).

As already we have already determined in Chapter Two, subject knowledge from the perspective of a teacher or student of GCSE English can be seen to exist in two distinct forms: explicit and tacit. Moreover, we can recognise that the relationship between these two types of knowledge are complex and interlinked. To exemplify, explicit knowledge tells us that a metaphor is a linguistic device in which a word or phrase is applied to an object or action to which it is not literally applicable; tacit knowledge informs us of how metaphor might be used in speech or in writing in specific contexts to elicit a reaction in our audience dependant on context, be it thought-provoking, humorous, pedagogical or otherwise. In the above example, it is conceivable that the explicit knowledge of what a metaphor is can be codified and communicated between parties. Ultimately, this is what makes it explicit knowledge. The same cannot necessarily be said for the example of tacit knowledge. Even with an understanding of what a metaphor *is*, it is impossible to codify how it might be used, and what appropriate entailments such a metaphor might apply to, in any given situation. The conditions for use in such situations are realised *in the moment* and are shaped by contextually complex social, environmental and linguistic factors that render a prescriptive and absolutist view of this knowledge as redundant.

It is in this example above, and countless others like it that draw distinctions between explicit and tacit knowledge, that we can locate the dissolving of boundaries between ontology and epistemology that Waring argues for above. To recognise tacit knowledge as an authentic source of meaning that underpins our understanding of reality is to subscribe to an ontology that rejects positivist ideals. This enquiry sets out to explore in depth concepts including practice, skill and judgement, each of which have their foundations in tacit knowledge. In order to fully explicate these concepts, and others closely associated with them, it seems appropriate to attempt to adopt an interpretivist epistemology in conducting this research.

Sustained throughout any discussion of an interpretivist epistemology is the role of social practice, interaction and co-creation. Research participants are of course of interest as creators here, but the role of the researcher themselves must to be considered, in that they too ascribe to the notion of interpretive epistemologies and are themselves bring their own values, beliefs, experiences and version of reality with them. This is particularly prominent in action-led practitioner research in which the research focus is located specifically within the practising domain of the researcher. Scott and Usher (2002) observe that “human action is given meaning by interpretive schemes or frameworks. It follows from this that as researchers [...] we too seek to make sense of what we are researching and we too do so through interpretive schemes or frameworks’ (2002:19). They go on to elaborate on what is referred to as the ‘double hermeneutic’, a term originally coined by Giddens (1982), that accounts for how in social research both the subject (the researcher) and object (other people)

of research have the same characteristics of being interpreters or sense-seekers (Scott and Usher, 2002:19).

As suggested above we can find a commonality in understanding that can be reached across multiple research agents, in the form of interpretive schemes or frameworks. As co-interpreters, these can help negotiate knowledge that is perspective-bound and partial, and relative to that framework (Scott and Usher, 2002:19). One example of a common framework is language, in that this provides humans with a vehicle through which we can articulate our experiences of the world with one another, and through dialogue reach what we perceive to be mutual understanding. Interestingly, the term hermeneutics has its disciplinary roots in interpretation of language, and historically has seen application as a methodology for interpreting meaning from biblical and philosophical texts. In modern applications of the term, this has broadened to account for the interpretation of meaning in all forms of language, including written and spoken. Underpinning this is what Zimmermann (2016) considers to be a key concept within the domain of modern hermeneutics:

> ‘Fusion of horizons: this [...] describes the nature of understanding as integrating what is unfamiliar to use into our own familiar context, so when we understand something we fuse someone else’s viewpoint with our own and in this encounter we are transformed because it broadens our mind.’ (Zimmermann, 2016)

This concept is helpful for two reasons. Firstly, it acknowledges how interpretivist frameworks, such as language, can act as mediators through which the sharing and co-creating of understanding can be achieved. In this, it provides a practical means of

application to the otherwise theoretical concept of interpretivism as an informant of a methodological approach in research design. Secondly, it elaborates on the notion of perspectival-bound and partial knowledge by explicitly referencing the role that other agents, those external to us, play in this.

In accordance with the above, the methodology adopted by an interpretivist researcher needs to be one that actively and comprehensively seeks to locate the perspectives of others, in an attempt to build towards a representative picture of the constituent parts and the whole of multiple versions of reality. Beyond methodology, we can look to specific methods of data collection that can help contribute towards this. In general terms, qualitative approaches offer individuals greater liberty in the manner and form of how they convey their experiences of the world when compared with quantitative methods. A more comprehensive discussion of the methods employed in this enquiry, including a justification for why these have been chosen is featured below; notwithstanding, as Grix (2002, 2010) upholds, there is a necessity in research for the ontological and epistemological disposition of the research and the researcher to inform the methodological approach that is most appropriate, which then informs the methods of data collection that best suit. If we are to adopt an interpretivist paradigm and operate in respect of concepts such as the ‘fusion of horizons’ then we must attempt to explicate our participants’ versions of reality as comprehensively as possible, minimising bias, assumption or neglect of potentially unforeseen matters.

In seeking to do this, there is a need to recognise that interpretations are not only perspectival and partial. Scott and Usher (2012) state that ‘as well as being

perspectival and partial, interpretations are always circular. The interpretation of part of something depends on an interpretation of the whole, but interpreting the whole depends on an interpretation of the parts' (2012:19). In this tradition we can recognise the notion of forming knowledge as one that exists not on a linear or cumulative scale, but as one that is circular, iterative and spiralling (2012:19).
Let us consider the example in which a teacher is assessing a student's piece of creative writing. The success of specific elements of creative writing can be determined through interpretation of micro-level analysis of the text; this might include subject-tense agreements, the correct spelling of words, the variance and intent of vocabulary employed. These constitute a form of partial knowledge - that in this example provides insight into conventions of grammar, spelling and meaning making - that help inform a teacher of how competent this piece is. But the success of a text is not alone determined by the application of, and judgement with, these specific partial forms of knowledge. The text as a whole must also be considered without breaking it down into its constituent parts, for this too represents an important line of interpretation that will inform how successful the creative writing piece is. The teacher might consider how the text *feels*, if it *flows,* or ask themselves "*does it talk to me?"*, that is to say, does it resonate with me as a human being. It is the negotiation between knowledge of the partial and whole that allows a teacher to interpret meaning from the text, and accordingly determine how successful it is. Ultimately, both the part and the whole are dependent on one another in order for an interpretation to occur.

It is here that we can locate the circularity of meaning making through interpretation. Scott and Usher (2012) uphold that the 'circularity of interpretation [...] always takes

place against a backdrop of assumptions and presuppositions, beliefs and practices, of which the subjects and objects of research are never fully aware and which can never be fully specified' (2012:19). Knowledge-forming through interpretation does not take place in a vacuum, but rather is bound by the context in which the interpreter inhabits. The 'assumptions, presuppositions, beliefs and practices' of the teacher are going to naturally impact on the meaning-making they are deriving when attempting to assess a student's creative writing. This poses something of a challenge when framing the practitioner-researcher (myself in this case) in that my own assumptions are of course (like the other subjects in this research) perspectival and subjective in nature. This poses something of a challenge when framing this teacher as a potential research subject, in that these assumptions and the like are themselves perspectival and subjective in nature. This challenge is exacerbated further by the need for the researcher to themselves interpret meaning from a subject who is interpreting meaning about the world. Research involves interpreting the actions of those who are themselves interpreters: it involves interpretations of interpretations - a double hermeneutic at work (2012:20).

In view of the above double hermeneutic, what is the best way for the interpretivist researcher to proceed? Gadamer (1975) offers one solution. He argues that it is impossible for researchers to escape from 'pre-understandings' but that this is not problematic. Rather, it is through these pre-understandings, far from them being prejudices or biases, are put to risk, tested and modified through the process of interpretation in the encounter with what one is trying to understand. To know, one must be aware of one's pre-understandings even though one cannot transcend them (Gadamer, 1975; Scott and Usher, 2012:20-21). It is here that we can refer back to

the concept of the ‘fusion of horizons.’ As a researcher’s own perspective cannot be put aside during an enquiry, knowledge is sought while grounded in this standpoint. With reference to another consideration of methods, Scott and Usher (2012) maintain that this ‘requires a dialogic situation, one where researchers are able to bring their pre-understandings into contact, through dialogue, with the pre-understandings of the researched and other researchers. However, the condition for this is that dialogue must be free and unconstrained by structural/ideological inequalities.’ (2012:24).

## Locating an ontology and epistemology for this enquiry

The discussions above chart how paradigms can inform how educational research is understood, designed and evaluated. Positivist paradigms offer absolutist views of the world in which knowledge and meaning are seen as objective reality. In contrast, interpretivist paradigms support the idea of locally constructed versions of reality, and that meaning and knowledge cannot be detached from the context in which they exist. Presented in the above section of this Chapter are some brief justifications for the adoption of an interpretivist paradigm in this enquiry. As presented in Chapter One, judgement of student creative writing is a personal activity. While external standards dictate the quality indicators that a teacher should be using when assessing, these standards are products of the social world and are to be understood by individuals who are participating in it. As such, the interpretation of these standards is a personal activity that takes place in a locally constructed version of reality. One aim of this research is to explore new kinds of knowledge that teachers can acquire through assessing through comparative judgement. Another is to explore

what learners' adaptive comparative judgement decisions tell us about their understanding of creative writing as a field of study within the discipline of English Language. As such, the intention of this research from an ontological position will be to better understand the world in which research participants (including myself) inhabit through "reporting multiple perspectives, identifying many factors involved in a situation, and generally sketching the larger picture that emerges" (Creswell, 2012:47).

Following from this alignment to an ontology of interpretivism, the epistemology that this research enquiry adopts conforms with accordant ideas of what knowledge is. Both explicit and tacit knowledge are viewed with equal validity as forms in which meaning about the world reside. Moreover, knowledge is seen as a socially-owned and constructed which is given form through the negotiation of interpretive frameworks by co-constructing agents. The idea of a 'fusion of horizons' provides me as a researcher one such framework, through which the experiences and perspectives of multiple agents can be exchanged and fused to ultimately lead to a transformation in understanding.

## Research quality: adequately representing the research context

Coe et al. (2017) argues that 'an even harder task than defining educational research is defining good research' (2017:12). While no universal criteria for determining what makes good research exist, he offers questions that might be used as a way of evaluating the quality of a piece of research. Coe et al. position these questions as

evaluative tools to be used after research has been conducted, but there is value in considering these questions when planning research too, in that they can act as a framing device to ensure that considerations in relation to specific quality markers are made. Some of the questions feature in subsequent discussions below. Alongside these questions resides an account of how I aim to address these points in this enquiry. These questions are interspersed throughout the following discussions that focus on research question construction, methods, participants, interpreting meaning and ethics, the first of which is seen below:

> *"How realistic or representative are the contexts in which the research was done? Are they described adequately?"* (Coe et al., 2072:13)

As this enquiry is practice-focused the context in which it is situated allows the opportunity to explore the research aims in an accurate representation of an authentic environment. To ensure that this authenticity is maintained, in designing this research I am mindful to ensure that the methods of data collection used are as genuine and non-contrived as is possible. The methods used align as closely as possible with common practices that teachers and learners already undertake on a regular basis, so as not to deviate from, and thus misrepresent, the context in which they are typically situated.

Chapter One of this thesis began with a short description outlining the institutional context in which I work, together with an account of how the perceived problem of inadequate assessment of GCSE English creative writing emerged in practice. This

represents a partial overview of the context in which the research is situated, but cannot be considered to be wholly so. As such, it is my intention in this enquiry to have this context more comprehensively built upon and made visible through the gathering and exploration of various voices, of teachers and learners, through the methods adopted in this study. While the context in which the enquiry is located is not a subject of this research directly, it frames all constituent elements of it. As such, matters of context are explored alongside those that feature as direct research aims. This, as Dornyei (2007) notes, aligns with an explicit goal of qualitative research: 'exploring the participants' views of the situation being studied'" (2007:38).

## Ethics

This enquiry is planned and designed with specific consideration towards participant consent. Liamputtong (2009) defines informed consent as the procedure to provide sufficient information to individuals to decide if they want to get involved in the study or not after being informed of the purpose of the research, research procedures, any potential risks and alternatives. In this enquiry I fully briefed all potential participants, both teachers and students, in advance of my conducting any research involving them. Emails were initially sent to teachers providing information as to the purpose and design of the research, enquiring if they would like to be involved. This was then followed by a meeting with each teacher who declared an interest in which further information about the research was shared, including what their specific role would be. At this juncture if teachers were happy to be involved they signed an Informed Consent Form. Teachers were explicitly told that they were entitled to opt out of the

research and to have their involvement not reported in any data at any point they wished.

For my students I initially shared a one-page information sheet about my research and what its aims were to give them some time to decide if they wished to participate. Following this, students that declared an interest in participating attended a group information session with other interested students in which they were briefed as to the design of the research and what their role in it would be. At this point students that were happy to participate in the research signed an Informed Consent Form. As above, students were explicitly told that they were entitled to opt out of the research and to have their involvement not reported in any data at any point they wished.

Another ethical issue concerned confidentiality, in view of which researchers have a responsibility to "ensure they do not disclose identifiable information about participants through various processes designed to anonymise them" (Wiles et al. 2006:3). This was paramount to this study particularly with regard to my working with teachers, as the data I was acquiring with regard to the assessment of student work could be construed as providing an insight into the quality of that teacher's ability to accurately assess student work, and by extension be understood as a measure of their performance as a teacher. The teachers that opted to participate in this study did so knowingly of this fact, but there is nonetheless a responsibility to protect the anonymity of each participant in respect of this. In appreciation of these ethical repercussions I assigned each teacher and student a neutral identifier (i.e., teacher A, student D) to replace their actual name in the study, so they had their identity protected.

Some of the data that was gathered during this study was obtained through interactions with teachers and students in a public domain and came about as a result of the social- constructivist environment in which the participants and researcher were located. It was important to recognise that by publishing the interactions in this enquiry the data presented is done so in a decontextualized manner, far removed from the environment, be that office or classroom, in which they were recorded. As the nature of research invites scrutiny into all available data it was essential to protect the identities of all of the students and teachers so that the interactions that feature in this enquiry are not attributable to any one individual, but still represent the authentic views of real teachers and students.

## Research methods

This section presents a justification for the methodology and methods employed in the thesis. It begins with a description of the sample of participants recruited for this research.

### The participants

This research involved both teachers and students at the college. This section provides a brief description of who these participants were and how they were engaged initially in the research process.

In total twelve different GCSE English teachers participated in the research. Across the sample there was a significant range in the amount of experience teachers had in teaching and assessing GCSE English, with some having taught the qualification for a number of years, and others having done so for a year or less. One teacher was a practicing GCSE English teacher who was currently undertaking his teacher training qualification. Teacher participants were recruited to the research at two separate intervals: seven were engaged in May 2018 and the remaining five in October 2018. All twelve participating members were unique, that is to say that no participants of the first seven engaged in May 2018 joined as one of the five teachers in October 2018. With the team GCSE English teaching team comprising approximately ten teachers at any given time, the engaging of research participants at two intervals across academic years (May 2018, during the 2017-18 academic year; October 2018, during the 2018-19 academic year) meant that teachers that joined the team in summer 2018 could participate, and a comprehensive sample was recruited. A more comprehensive profile of some of the teachers involved in the research can be found in Chapter Four.

The research was first introduced to prospective teacher participants through the same method. On both occasions members of the GCSE English teaching team were given an introduction to the research via a briefing at a team meeting. The 'Information sheet for prospective participants' (see appendix item 8.1) was shared during the briefing, and teachers who were interested in participating were encouraged to make contact via email to register their interest. After they had emailed me declaring an interest in participating, teachers read and, on agreeing to the terms, signed a research consent form (see appendix item 8.2) to formalise their

participation in the research. The terms included the ability for any participant to no opt out of the research at any point if they wished.

In total twenty-five students took part in this research as participants. All were students that I had taught GCSE English in the 2017-2018 academic year. I chose to initially engage students I had worked with here as I hoped the rapport I had developed with them working over the academic year would help encourage their participation in the research. I also hoped further into the research process that students that I was familiar with might be more forthcoming in sharing their experiences of the subject, particularly during the semi-structured interviews.

Potential students were initially engaged in this research by attending an initial research briefing that took place at the end of their timetabled GCSE English class with me. Attendance at this briefing was optional. The 'Information sheet for prospective participants' (see appendix item 8.1) was shared during the briefing, and student questions were answered. Those who were interested in participating were encouraged to let me know either in person or via email within two weeks. After students had declared their interest in participating I used email as the primary method of contact with them about the research. I did this as I was conscious of not making reference to the research during our timetabled class time, so as not to impact on the students that had opted not to participate. Another short session was then held for the student participants that had opted to participate, in which they read and, on agreeing to the terms, signed a research consent form (see appendix item 8.2) to formalise their participation in the research. The terms included the ability for

any participant to no opt out of the research at any point if they wished, and this was made explicit to students.

The twenty-five students that participated comprised of two classes: sixteen were students aged between 16-18 on a full-time study programme; the remaining nine were all adult (19+) learners on a part-time GCSE English programme. The motivation, previous experiences and disposition towards studying GCSE English varied considerably across both student groups. For the students between the ages of 16-18, GCSE English was a compulsory qualification they were required to study as part of their enrolment on a study programme, the main composition of which was a vocational or academic qualification that they had chosen to study. Their enrolment on a GCSE English qualification at the beginning of the academic year was automatic because they had each previously achieved a grade '3' in GCSE English, that is to say a grade just under the 'pass' threshold. For these students, the prospect of revisiting a curriculum they had already studied, in some instances more than three times prior, in an attempt to pass a qualification they had recently 'failed', posed significant challenges to their perceptions, of both the subject and their ability as students of the English Language.

For the adult students that made up the remaining nine participants in this research, motivations for studying GCSE English varied. Several required a 'pass' standard grade in GCSE English in order to access higher education and professional qualifications at university. Other students had opted to study GCSE English having previously completed prior levels of study, initially in English Speakers of Other Languages (ESOL) and then Functional Skills qualifications. For these students

GCSE English represented the next step available to them in continuing their study of English. The nine adult students represented a widely diverse cohort: their ages spanned from those in their early twenties to mid-sixties; seven students had been raised and educated in a different country up to the age of fifteen; and six of those did not speak English as their native language. The considerable diversity in the adult student participant group, in addition to the students in the 16-18 group, represents a diverse student cohort that can be considered representative of the wider GCSE English student cohort that studies at the college.

A profile of the student participants involved in this research can be found below. This includes their age category, whether they opted to participate in the semi-structured interviews, their ethnicity, if they are entitled to a bursary, how many attempts they have previously made at achieving a grade 'C' / grade '4' in GCSE English, and if they had any declared learning support needs.

| Student | Age group | Interviewed | Ethnicity | Entitled to bursary | GCSE English attempts | Learning support needs |
|---|---|---|---|---|---|---|
| **A** | 16-18 | Yes | White British | No | 1 prior (school) | None declared |
| **B** | 16-18 | Yes | Asian British | No | 1 prior (school) | None declared |
| **C** | 16-18 | Yes | Black British | No | 3 prior (1 at school, 2 at college) | None declared |
| **D** | 16-18 | Yes | White British | No | 1 prior (school) | None declared |
| **E** | 16-18 | Yes | Asian British | Yes | 1 prior (school) | None declared |
| **F** | Adult | Yes | African | N/A | No prior attempt made | None declared |
| **G** | 16-18 | Yes | Asian British | Yes | | None declared |
| **H** | 16-18 | Yes | White British | No | | None declared |

| | | | | | | |
|---|---|---|---|---|---|---|
| **I** | Adult | Yes | Caribbean | N/A | No prior attempt made | None declared |
| **J** | Adult | Yes | White European | N/A | No prior attempt made | None declared |
| **K** | 16-18 | Yes | White British | No | | None declared |
| **L** | 16-18 | Yes | White British | No | | None declared |
| **M** | Adult | Yes | White British | N/A | 1 prior (O-Level) | None declared |
| **N** | 16-18 | No | Asian British | No | | None declared |
| **O** | 16-18 | No | Black British | Yes | | None declared |
| **P** | Adult | No | White British | N/A | No prior attempt made | None declared |
| **Q** | 16-18 | No | White British | Yes | | None declared |
| **R** | 16-18 | No | Black British | No | | None declared |
| **S** | 16-18 | No | Asian British | No | | None declared |
| **T** | 16-18 | No | Black British | No | | None declared |
| **U** | 16-18 | No | White British | No | | None declared |
| **V** | Adult | No | Irish | N/A | No prior attempt made | None declared |
| **W** | Adult | No | Chinese | N/A | No prior attempt made | None declared |
| **X** | Adult | No | Asian British | N/A | No prior attempt made | None declared |
| **Y** | Adult | No | Black British | N/A | 1 prior (school) | None declared |

## The research methods

Qualitative research typically works with small sample sizes, which are selected purposefully to permit in-depth inquiry into, and understanding of, the phenomenon concerned (Patton, 2002:45). Kincheloe (2012) observes that selecting appropriate methods for research is an activity often steeped in ideological bias, noting that "We come to recognise that there are no value-free, privileged knowers who ask ideologically unfettered questions about the methods they will employ in their studies" (2012:216). This point is worthy of consideration, as research is often conducted with the purpose of yielding new knowledge that can be harnessed, and as such methodologies, findings and analysis can be skewed to favour recommendations that are quantified, tangible and readily actionable.

Quantitative based educational research, particularly of the state-funded variety, is currently *de rigueur*. But, as Tobin (2006) argues, such traditions should be challenged: "No matter how much the mavens of evidence-based inquiry in right-wing movements may insist that there is one right way to produce educational research, we are convinced of the power of multiple ways of seeing the world - the educational world in particular" (2006:1). We can look again to Wiliam's (2019) 'complications' of using positivist paradigms in educational research as having shortcomings. The 2016 Government White Paper for education, Educational Excellence Everywhere, acknowledges that, "It is not yet easy as it should be for teachers to find and use evidence to improve their teaching practice because the evidence base is patchy, difficult to access or to translate into action" (2016:39). The challenge that is faced by teachers is that educational research evidence can sometimes infer, either explicitly, implicitly or without intending to, that results

observed in one research context are replicable across a multitude of others. Quantitative methods of research, and the positivist paradigms that they are often found in proximity to, can be considered at least partially responsible for this challenge.

This research seeks to explore human experience and unearth new knowledge and understandings about assessment practice in GCSE English creative writing. It intends to present practicable findings that can be accessed and considered by other practitioners in similar and different contexts, to find meaning relative to their experiences and in reference to their own practice. In consideration of this, the opportunity to qualitatively investigate my colleague's practice, and the experiences of my colleagues and students in my own context, represents one possible way to offer evidence-based insights into these practices practice for others. To achieve this, the research methods have been selected with an appreciation of the characteristics of qualitative inquiry.

Table 1 below provides an overview of the methods used in this enquiry. It provides details of the participants, how each method aligns to the research questions posed, and the form and nature of the data being gathered. A more comprehensive discussion of each method is available below this table.

| Method | Method description | Participants | Alignment to research questions | Data analysis |
|---|---|---|---|---|
| **Method 1** | **Adaptive Comparative Judgement trial** (comprising a workshop and subsequent individual judging) | Seven members of the college's GCSE English team | RQ1<br>Sub RQ2 | ACJ assessment decisions gathered from NoMoreMarking software. Testing reliability, accuracy of judgements |
| **Method 2** | **Adaptive Comparative Judgement workshop** | Five members of the college's GCSE English team | RQ1<br>Sub RQ2 | ACJ assessment decisions gathered from NoMoreMarking software. Testing reliability, accuracy, time taken per judgement. |
| **Method 3** | **Semi-structured interviews** | Five members of the college's GCSE English team | RQ1<br>Sub RQ1<br>Sub RQ2 | Semi-structured interviews, building individual teacher profiles, gaining perspectives into use of ACJ for assessment of creative writing |
| **Method 4** | **Student questionnaire** | Ten GCSE English students | RQ1<br>Sub RQ3 | Student feedback on use of ACJ, focusing on the perceived value of ACJ as a method of peer learning (1), ACJ as helping to develop an understanding of the subject (2), and if it was an effective use of time (3). |
| **Method 5** | **Semi-structured interviews** | Thirteen GCSE English students | RQ1<br>Sub RQ3 | Semi-structured interviews, gaining student perspectives into use of ACJ for the peer assessment of creative writing |

*Table 1: an overview of the research methods used in this interpretivist synthesis*

## Method 1: Adaptive Comparative Judgement trial (workshop and subsequent individual judging)

The first research method selected was an adaptive comparative judgement trial that involved seven members of the college's GCSE English team. This was the first trial of using ACJ approaches to the assessment of creative writing at the college. It spanned two weeks in May 2018. The trial comprised each teacher being allocated 110 comparative judgement decisions across a set of eleven student creative writing scripts (110 being the number of judgements recommended by NoMoreMarking to ensure a reliable result). All judgements were completed using NoMoreMarking's Adaptive Comparative Judgement online software.

### Method 1: The student creative writing scripts

Each script was an authentic item of work written by a different student. They had been written in a classroom setting under exam conditions. The task itself was similar in its design to an AQA GCSE English creative writing question, but had been created for the purposes of this study to ensure that students were not answering a question, or that teachers were not judging scripts, that they had seen before. The task students completed can be found in the appendix, under the label *appendix item 8.4.*

### Method 1: The adaptive comparative judgement workshops

Teachers involved in the trial were invited to three comparative judgement workshops, spanning across a week in mid-May 2018. These three workshops took place at different times and days to allow them to fit in around teachers' timetables. In order to take part in the trial teachers had to attend at least one workshop;

attendance at more than workshop was optional. All seven teachers attended at least one workshop session, with two attending all three. Each workshop session ran for 90 minutes. Each was structured the same way: the first twenty minutes comprised an overview of comparative judgement and an introduction to the ACJ software; the remaining time was then allocated to teachers completing their allocated number of judgements.

When judging, teachers were asked to choose the 'most proficient' text from each combination. They were given no further instruction. Other research into ACJ approaches to assessment have explored providing alternative forms of stimulus and instruction; Whitehouse and Pollitt (2012) employ an 'importance statement', which comprises the aims of a specification that 'established a link between the rigour of the [...] specification and making holistic judgements without reference to a mark scheme' (2012:6). Although use of an importance statement might have proven useful for teachers when forming judgements, one aim of this research is to explore how successful teachers could make assessment decisions in lieu of supporting material. In the example above there is a risk that the importance statement might interfere with the teachers' ability to form a holistic judgement, even if it is not as detailed or prescriptive as a mark scheme. As such, in this trial no mark scheme, assessment standards, or any other materials were shared with the teachers to help in reaching these judgements. All judging took place individually, with each teacher working at their own station. All of those involved were asked to avoid discussing decisions they were making so as not to interfere with the judgements that others were making.

From a purely practical position, workshops were not necessary for the successful running of this ACJ trial. This is as all judging can be completed remotely using NoMoreMarking's judging software, providing the judge has an internet enabled device and the link to access the sample. Despite this the decision to arrange and invite teachers to workshops was a deliberate one. What they ensured was that teachers interested in the project were sufficiently briefed and inducted as to what ACJ is, the purpose of the task, and the research as a whole. They also represented a designated time and provided a location in which teachers participating in the research could undertake their judgement of allocated scripts. This was crucial, in that each of the teachers engaged in the trial were doing so of their own choosing and electing to give up time designated for planning and preparation. It was my intention that these workshops would provide motivation for the teachers involved, if they could see their other colleagues participating too.

**Method 1: subsequent individual judging**

Following the workshops, the teachers were given a week in which they could complete any outstanding judgement decisions they still had remaining to make from their allocation. This was to be done remotely. Teachers were reminded again not to refer to any mark schemes or assessment criteria, and instead consider which of the two was the 'most proficient'. While these conditions were not as controlled as with the workshop sessions I felt it important to allow teachers the opportunity to complete their judgements where and when they chose so as to reflect a more authentic way in which comparative judgement might be implemented in future instances.

**Method 1: alignment to research questions**

In respect of the research questions framing this study, this method was aligned to providing insight to answer the following questions:

***RQ1:*** *What are the benefits and challenges of using adaptive comparative judgement approaches when assessing GCSE English creative writing scripts in a Further Education institution?*

***Sub-RQ2:*** *How can adaptive comparative judgement be used across a team of teachers to standardise assessment practices?*

## Method 2: Adaptive Comparative Judgement workshop

The second research method was an adaptive comparative judgement workshop conducted with five of the college's GCSE English teachers. These five individuals were all different to the seven that had engaged with the initial ACJ trial. I had intentionally focused on engaging these individuals when organising this workshop as I wanted a broad sample of participation from across the team. This workshop took place in October 2018 and ran for 90 minutes in duration.

**Method 2: The student creative writing scripts**

One of the intentions with this workshop was to mimic conditions that would be found in a typical standardisation activity, namely, teachers contributing their student's work to the sample that was being considered and standardised. In preparation for the

workshop, all participating teachers agreed the task their classes would complete in order to contribute to the sample. This was the same task that had been completed by students for the previous ACJ trial *(appendix item 8.4).* In the weeks preceding the workshop each teacher had one of their classes (comprising at least twelve students) complete the task under exam-style conditions, and submit these scripts to be entered into the sample. The pieces were not seen or marked by the teachers before being shared.

**Method 2: selecting the student creative writing scripts**

In total seventy-six scripts across the five classes were submitted in the weeks preceding the ACJ workshop. After receiving these seventy-six scripts each was given a unique reference number. As the total sample size for this ACJ activity was to be only fifteen, it allowed scripts from students across all five classes to feature in the final sample. This was valuable, as it meant that a proportional number of scripts from each class could be selected, and teachers would be judging work that was not solely completed by their students and that they would be exposed to other students' creative writing.

In order to select the fifteen scripts that would make up the final ACJ sample a random number generator was used to select three scripts from each teacher's sample, using the unique reference numbers given to each script to facilitate this. To exemplify, teachers 1's submitted sample comprised fourteen student scripts. Accordingly, each script in this sample was assigned a unique reference number between 1-14. From this sample scripts 5, 12 and 13 were selected by a random number generator (1-14) to go into the main sample. This same pattern was adopted

for the other sets of scripts that had been submitted by each of the other four teachers. In this configuration each teacher had three of their students' scripts contribute to the overall sample of fifteen. After these texts had been selected, they were digitally scanned, and the student author's name was removed to help prevent possible teacher bias when judging.

**Method 2: The adaptive comparative judgement workshop**

This workshop was 90 minutes in duration. As with method 1's ACJ workshops, the first twenty minutes comprised an overview of comparative judgement and an introduction to the ACJ software, and the remaining time was allocated to teachers completing their allocated number of judgements. The total number of judgements was set at seventy, the number recommended by NoMoreMarking to ensure a sufficient number of judgements per script (at an average of 23 decisions per script).

Each teacher undertook adaptive comparative judgement individually at their own station, and worked to complete as many comparative judgements as they could in one hour from the sample of fifteen texts. This differed to method 1, in that in this workshop a time limit was imposed. After the one hour of allocated judging time no more judgements were to take place. This was not necessarily to test the speed at which judgements were being arrived at (although this is reported through the use of the NoMoreMarking software, and will be briefly discussed in Chapter Five). Rather, this was to gauge how many judgements were being made per hour across each of the judges, with the intention to use this as a means of comparison against (i) the other judges, and (ii) the accuracy of their own judging decisions.

As in method 1, teachers were asked to choose the 'most proficient' text from each combination, and no mark scheme, assessment standards or supporting documents were shared with them to help in reaching these judgements. Similarly, all of those involved were asked to avoid discussing decisions they were making so as not to interfere with the judgements that others were making.

**Method 2: alignment to research questions**

In respect of the research questions framing this study, this method was aligned to providing insight to answer the following questions:

***RQ1:*** *What are the benefits and challenges of using adaptive comparative judgement approaches when assessing GCSE English creative writing scripts in a Further Education institution?*

***Sub-RQ2:*** *How can adaptive comparative judgement be used across a team of teachers to standardise assessment practices?*

## Method 3: Semi-structured interviews with teachers

These interviews were conducted with the five GCSE English teachers who participated in the ACJ workshop. They all took place within a week of the ACJ workshop that formed *method 2*. All interviews were conducted one-to-one. The intention of these interviews was, as Dornyei (2007) notes to 'explor[e] the participants' views of the situation being studied' (2007:38), and to better understand

some of the complex issues of tacit knowledge and individual judgement practice through 'reporting multiple perspectives, identifying many factors involved in a situation, and generally sketching the larger picture that emerges' (Creswell, 2012:39).

Each interview comprised three sections, with questions following a distinct theme in each of these.

The first section of questions focused on ***teachers' experiences and training in teaching and assessing GCSE English.***

*1. How many years have you taught GCSE English in a Further Education setting?*

*2. What formal training, if any, have you participated in teaching and assessing GCSE English? How effective was this?*

*3. What informal training, if any, have you participated in teaching and assessing GCSE English? How effective was this?*

These first three questions sought to establish background information about the teacher. The focus on training in questions 2 and 3, and the use of both 'formal' and 'informal' qualifiers to describe any training they might have undertaken, was an intentional distinction. The aim here was to identify what each of the teachers interviewed, felt represented formal or informal training in respect of teaching and assessing GCSE English. The answers to these three questions helped establish a

profile for each teacher judge that gave representation to their experience of teaching GCSE English. These profiles are presented in Chapter Five.

The second section of questions focused on **reflecting on the use of adaptive comparative judgement:**

*4. What is your experience of assessing creative writing through adaptive comparative judgement?*

*5. Did this approach to assessment change the way you viewed each script?*

*6. What have you gained through assessing with comparative judgement?*

*7. Do you have any other comments you'd like to make with reference to adaptive comparative judgement?*

These four questions set out to learn more about teachers' experiences of using adaptive comparative judgement as an approach to assessing creative writing. They were constructed in a way that conforms to the paradigmatic alignments discussed in the above section on epistemology and ontology, in that they intend to gain insight into different perspectives from multiple agents. Each of the questions were open and encouraged the interviewee to share detailed responses. In instances where teachers shared less detailed responses, prompts were used to encourage additional reflections and contributions from them.

Each interview took place on a one-to-one basis, which allowed each teacher to share their personal experiences of using ACJ. One-to-one interviews were chosen intentionally to prevent any dilution or interference of ideas between teachers, something that might have occurred if focus groups were adopted as a method of data capture. In doing this, it was hoped that it would be possible to sketch together common themes that were identified through these one-on-one interviews, and report them as significant due to this commonality with at least some degree of confidence. These common themes will be identified in Chapter Five, and expanded on in Chapter Six.

The third section of questions focused on **the practice of undertaking adaptive comparative judgement:**

*8. Which script is more proficient as a piece of creative writing?*

*9. Describe what is helping you make this judgement? What are you drawing on?*

In the last part of the interview teachers were introduced to two creative writing scripts and asked to narrate the thinking they were undertaking in comparatively judging these two scripts in detail. These scripts were paper-based, and teachers were given ample time to read them both before being posed the questions above. The two scripts selected had been ranked as being similar in proficiency as determined by the judgements formed in method 2's ACJ workshop. The intention here was to pose the teachers a comparative decision that was not easily

immediately resolvable, and as such gain insight into the process they were undertaking in identifying greater proficiency.

**Method 3: alignment to research questions**

In respect of the research questions framing this study, this method was aligned to providing insight to answer the following questions:

***RQ1:*** *What are the benefits and challenges of using adaptive comparative judgement approaches when assessing GCSE English creative writing scripts in a Further Education institution?*

***Sub-RQ1:*** *What new knowledge can be acquired by teachers as a result of undertaking adaptive comparative judgement and what function does this serve teachers of GCSE English in an FE context?*

***Sub-RQ2:*** *How can adaptive comparative judgement be used across a team of teachers to standardise assessment practices?*

## Method 4: Student questionnaire;

## Method 5: Semi-structured interviews with students:

Methods 4 and 5 address the important matter of students using adaptive comparative judgement to peer assess the quality of their peers' creative writing. The

goal here was not to check the quality or reliability of the students' assessment decisions that were being arrived at, as with the teachers in method 1 and 2. Rather, these methods sought to gain insight into students' experiences and reflections on using this approach to assessment, more in line with the method adopted with teachers in method 3.

**Methods 4 & 5: The student participants**

As discussed above, students that took part in completing these questionnaires and interviews were students that I had taught GCSE English in the 2017-2018 academic year. Students were invited to an adaptive comparative judgement workshop session in May 2018 that took place directly after their typical timetabled lesson. This took place at the end of the day, so did not clash with other timetabled sessions that students might have had. Students were fully briefed as to the goals of this research, and what the workshop would entail when the invite was shared. Attendance and participation in the workshop were optional, and students were informed that they could opt out or leave at any point if they wished.

**Methods 4 & 5: The student creative writing scripts**

The sample of creative writing scripts used for this adaptive comparative judgement workshop was the same as used in Method 2. I had considered creating a new sample of creative writing for this workshop using scripts that students in this class had written, but decided against it on account of potential personal biases that students might have in favour of, or in opposition to, scripts that they had personally written. For this workshop, all scripts were written by students different to those who

were undertaking ACJ. There were no names on the scripts so anonymity of each author was preserved.

**Methods 4 & 5: The adaptive comparative judgement workshop**

In total, twenty-five students attended the adaptive comparative workshop. The workshop lasted for two hours. In a similar construction to the workshops with teachers in methods 1 & 2, it began with an overview of what comparative judgement is. Students were given fifteen minutes to practice using the NoMoreMarking software to judge practice texts, following which all students said they felt comfortable with using the software.

Before students began comparative judging, they were given the explicit instruction that they were to be choosing “the better text” of the two in each pair. They were encouraged to reflect on what they felt was meant by the “better text”, and were given no additional supporting documents to help make their choices. Students then had forty minutes to complete as many comparative judgements as they could, using NoMoreMarking to judge the fifteen creative writing scripts that comprised the sample.

On completion of the ACJ activity, students were invited to share their experiences of using the approach. Students could either complete a digital questionnaire, participate in a semi-structured interview, or leave if they wished. Of the twenty-five that attended the workshop, two chose to leave after the ACJ activity, ten chose to complete the questionnaire, and thirteen opted to take part in a semi-structured interview.

**Methods 4: the student questionnaire**

I opted to include a questionnaire for students as I felt that it would provide a method for some students to share their experiences of using ACJ without having to participate in an interview, which some may have found daunting or uncomfortable.

Questions featured in the questionnaire were:

1. I have learnt from reading what my peers submitted.

   *Strongly disagree  Disagree  Neutral  Agree  Strongly agree*

2. This activity has helped me to understand what markers are looking for.

   *Strongly disagree  Disagree  Neutral  Agree  Strongly agree*

3. This activity was a good use of my time.

   *Strongly disagree  Disagree  Neutral  Agree  Strongly agree*

4. Please add any other comments you have about the comparative judgement process.

   *(free text response)*

The use of a Likert scale in questions 1-3 here was to help students share their experiences in reference to key themes, namely: the perceived value of ACJ as a method of peer learning (1), ACJ as helping to develop an understanding of the subject (2), and if it was an effective use of time (3). Perhaps most important was the

inclusion of a free-text box, which presented students with an opportunity to share their personal experiences. As students that opted to complete the questionnaire did not participate in the semi-structured interviews, this free-text box was crucial in gaining insight into student perspectives.

## Method 5: the semi-structured interviews

As noted above, thirteen students opted to participate in semi-structured interviews following the ACJ trial. One of the challenges in conducting interviews for research purposes is what Denscombe (2010:178) calls the “interviewer effect”. I was aware of the potential impact of this phenomenon, particularly in this instance, as I was working with students. In order to try and avoid this, I ensured that the interviews were open-ended and yielded natural dialogue but that maintained a consistency in the same basic information that was discussed (Bryman, 2004). I encouraged students to choose what configuration they would prefer for their interview, and whether they would prefer to be interviewed one-to-one or in a group. The configurations were decided by students as follows:

- Student D chose to be interviewed one-to-one
- Student L chose to be interviewed one-to-one
- Students K & M chose to be interviewed as a pair
- Students B, C, G & H chose to be interviewed as a group of four
- Students A, E, F, I & J chose to be interviewed as a group of five

Following from the above, five interviews took place in total. In each interview I had questions I wanted to ask, but similarly was mindful of losing any organic discussion points that might have arisen over the course of the interview. Bewley and Smardon (2007) note the value of effective dialogue for learning, stating "there is significant evidence in the student perception data that students value opportunities to talk about their thinking and learning and that through talking with others metacognition and flexibility of thinking is impacted on" (2007:7), and during the interviews I attempted to encourage mutual dialogue between students and myself so as not to restrict the possibility of interesting or unanticipated ideas of themes being shared.

These interviews were structured around the following two questions:

1. How did you decide what the better piece of writing was?

2. What helped you decide?

**Methods 4 & 5: alignment to research questions**

In respect of the research questions framing this study, this method was aligned to providing insight to answer the following questions:

***RQ1:*** *What are the benefits and challenges of using adaptive comparative judgement approaches when assessing GCSE English creative writing scripts in a Further Education institution?*

***Sub-RQ3:*** *What can learners' adaptive comparative judgement decisions tell us about their understanding of creative writing as a field of study in the discipline of English Language, and what are the subsequent pedagogical implications that follow from this?*

## Data analysis

In this enquiry it is important that all data gathered through the methods presented above, are considered to be pertinent, authentic and valuable in providing insights into the issue under examination. Discounting specific data from further analysis for any reason would jeopardise the integrity and trustworthiness of the research. As such, it is crucial to justify the approach taken in the selecting and analysing of specific data that is presented in Chapter Four: Findings, and the approach to analysis of this data featured in Chapter Five: Discussion.

### Research methods 1 & 2 – the adaptive comparative judgement workshops

Research methods one and two, present data gathered from the conducting of adaptive comparative judgement workshops with GCSE English teachers. This data are presented in the table format, taken directly from the NoMoreMarking software that was used to facilitate the adaptive comparative judgement work that teachers completed. The data presented in methods 1 and 2 represents all of the data gathered using these methods. All gathered information relating to the assessment

practice of the seven teachers that took part in the first workshop is reported in method 1, and of all five teachers that took part in the second workshop in method 2.

The NoMoreMarking software permits valuable insights into the assessment decisions that teachers make when comparatively judging respective quality across a sample of student creative writing scripts. This is gathered in the form of three separate pieces of information: the judge's infit score (1), their local score (2), and median time (3).

**Infit:** This metric represents the level of agreement between judges on scripts, with respect to the overall quality of the scripts that teachers are judging. Agreement is calculated through the NoMoreMarking software and uses Scale Separation Reliability (SSR) as a measure. In this system a lower score (1.0 or lower) is high agreement, representing little disagreement amongst judges between scripts. A score between 1.0 - 1.3 indicates 'some inconsistency' and a score in excess of 1.3 represents 'inconsistent' judging decisions in view of the other judges' decisions (NoMoreMarking, 2019).

**Local:** This metric represents how many judgements each teacher made in one hour.

**Median time:** This metric shows the duration of time spent judging each script individually.

Chapter Four reports on the findings from methods 1 and 2 report against each of these elements for all teachers that participated. What follows this is a deeper

analysis of what this information tells us about these teachers' assessment practices. This is achieved through comparison across the teachers. The quantitative nature of this data means that this comparison is possible with relative ease. Trends in this data and significant outliers are examined in further detail here.

## Research methods 3, 4 & 5 – analysis of qualitative data

Research methods 3, 4 and 5 comprise of data gathered through semi-structured interviews. In opting to use this method of data capture, it is important to recognise and pre-empt possible challenges that might be encountered when approaching analysis. In total ten interviews were conducted for this research featuring both teachers and students. All interviews lasted at least 20 minutes, and three lasted over 30 minutes. It is important to recognise that the process of using interviews as a method of data capture invariably leads to a significant amount of data being gathered. In practical terms it is impossible to report on and analyse every single utterance from each interview. Such problems are commonplace when adopting qualitative approaches to data collection in research. Nowell et al. (2017) note that:

> 'to be accepted as trustworthy, qualitative researchers must demonstrate that data analysis has been conducted in a precise, consistent, and exhaustive manner through recording, systematizing, and disclosing the methods of analysis with enough detail to enable the reader to determine whether the process is credible' (2017:1).

Accordingly, it is critical that a suitable approach to the analysis of all data gathered in these interviews is adopted. Moreover, this approach needs to be fully articulated so that consideration can be made towards the research credibility.

**Thematic analysis approach: trustworthiness**

This research uses a qualitative research dimension to generate knowledge grounded in human experience (Sandelowski, 2004). In order to achieve this a thematic approach to the analysis of qualitative data is used. This is a method for identifying, analysing, organizing, describing, and reporting themes found within a data set, and can help produce trustworthy and insightful findings (Braun & Clarke, 2006). In order to ensure that the thematic analysis of data is rigorous, Lincoln and Guba (1985) propose several criteria that strengthen the trustworthiness of the process.

First criterion advocated here is credibility. This is defined as the “fit” between respondents’ views and the researcher’s representation of them (Tobin & Begley, 2004). It is suggested that debriefing participants and sharing findings and interpretations with them after data has been collected can be useful in ensuring that views between researchers and their participants are aligned (Nowell et al.,2017). The credibility of this research has been promoted by participants receiving transcript copies of their interviews as well as an invitation to make amendments or corrections to anything they shared during interview if they feel their views were misrepresented in any way. Changes made to the interview transcripts are considered as

representative of the participants' views, and data that was removed by the participant is not reported on or analysed.

The second criterion of trustworthiness is transferability, and relates to the generalisability of the inquiry (ibid:3). This is achievable through providing comprehensive descriptions of the research through which transferral of findings to other contexts is achievable. This research achieves transferability by offering detailed accounts of important elements of the research, including the situated context (Chapter One), the research aims and methods (Chapter Three), the descriptions of participants (Chapter Three) and approach to analysis of data, as illustrated in this section. It is through this level of description that those who seek to transfer the findings to their own site can judge transferability (Lincoln & Guba, 1985).

Further criteria of trustworthiness relate to dependability and the use of audit trails. This is concerned with ensuring the research process is logical, clearly documented and that a clear rationale for decisions is present (Koch, 1994; Tobin & Begley, 2004; Nowell et al., 2017). In order to secure dependability in this research, audit trails are used to capture and record the processes of data collection and analysis that have taken place . Appendix item 8.5 '*Data collection methods summary*' comprises evidence of early planning in relation to the methods used in this research. This includes the questions that were asked during interviews with participants. Appendix items 8.7 '*Audio recordings from teacher interviews'* and 8.8 '*Audio recording from student interviews'* comprise of full audio recordings of all interviews that took place

with participants in this research. Full details relating to these auditable trails of the research process are available in the Appendices section of this thesis.

**Thematic analysis approach: phases of the process**

Nowell et al. (2017) outline a procedure for conducting thematic analysis that aims to meet the trustworthiness criteria outlined by Lincoln and Guba (1985). This comprises:

- Phase 1: familiarising yourself with your data
- Phase 2: generating initial codes
- Phase 3: searching for themes
- Phase 4: reviewing themes
- Phase 5: defining and naming themes
- Phase 6: producing the report

*(*adapted from Nowell et al., 2017*)*

The following section provides further details of each phase in this process, and explains how this research aligns with each phase.

**Phase 1:** familiarising yourself with data

Before coding and deeper analysis of qualitative data can take place, it is recommended that researchers read through their entire data set at least once (Nowell et al, 2017). In this research interviews with participants took place across

different intervals. This made this first phase of familiarising oneself with all data before coding began challenging, as a wait was required before all data had been gathered and an entire reading of the data could take place. In the very first instance this phase involved listening to the audio recordings of interviews and transcribing these into written format. Transcription was done as faithfully as possible, using audio playback software to pause, re-listen and slow down specific passages to ensure that these were captured accurately. An example of a transcription of a student interview can be found in the appendices section of this thesis, titled '*Appendix item 8.9 – student interview transcription excerpt'*.

**Phase 2:** generating initial codes

The second phase in data analysis is concerned with having ideas about what is in the data and thinking carefully about and identifying what is interesting about these ideas (Braun & Clarke, 2006). This phase of data analysis is reliant on the researcher being familiar with the entire data set. The aim here is to move from unstructured data to the development of ideas about what is going on in the data (Morse & Richards, 2002) through use of codes. When effective, these codes can capture the qualitative richness of the phenomenon under investigation (Boyatzis, 1998).

In this research, coding took place on the written transcripts of the participant interviews. The use of a coding framework that offered suitability and practical application was achieved by the use of a coding system tailored to each set of questions in the interviews. This coding system was flexible enough to be applicable

across both teachers and students. To demonstrate, responses to questions 1 and 2 in the student interviews shared a code with responses to questions 8 and 9 in the teacher interviews, in that both sets of questions were centred on how they arrived at their comparative assessment judgement and were concerned with identifying markers of good quality creative writing. An example of this coding system applied to a section of one student's interview transcript can be seen in the appendices section of this thesis, titled '*Appendix item 8.10 – coding of student interview excerpt'*

**Phases 3, 4 and 5:** searching for, reviewing and naming themes

In the context of thematic analysis, a theme is an 'abstract entity that brings meaning and identity to a recurrent experience and its variant manifestations. As such, a theme captures and unifies the nature or basis of the experience into a meaningful whole' (DeSantis and Ugarriza, 2000:362). Braun & Clarke (2006) remind us that themes are not dependent upon how many times something has been mentioned, but whether it captures something important in relation to the overall research question.

In this research the focus of interviews is wide-ranging. In the teacher interviews focus is placed on their professional development and experience in teaching and assessing GCSE English, their impressions of using comparative judgement, and on their impressions of good quality creative writing. The latter of these three is the sole focus in the student interviews. These different areas of focus meant that identifying emerging themes from the interviews was manageable, in that each offered distinct thematic categories aligned to that set of questions.

The emerging themes traced through examination of the qualitative data from the teacher and student interviews are discussed in more detail in Chapter Four. In this section, discussions include references to commonly occurring themes evident across multiple participants, and both teachers and students. Also included here are considerations towards themes that while pertinent and interesting were considered to not have enough data significance in the data to support them.

**Phase 6:** producing the report

King (2004) suggests that direct quotes from participants are an essential part of any final research report. In accordance with this, Chapters Four, Five and Six all include direct quotes from participants to illustrate the themes evident across data sets, and enrich the discussions that follow. The direct quotes are included often in isolation from the wider discussion that took place between the participant and the researcher, so contextual statements have been added when introducing these quotations to ensure authenticity and transparency in what is being reported.

# Chapter 4: Findings

This Chapter presents the findings from a range of methods undertaken as part of this enquiry. These include findings from the comparative judgement workshops conducted with GCSE English teachers, from semi-structured interviews conducted with the same teachers after they had used comparative judgement to assess learner creative writing scripts, and from interviews with students after they had used comparative judgement to peer assess creative writing scripts. Findings are presented in sections relating to each of the methods employed in this enquiry. Emerging themes and trends are highlighted in this Chapter, and are discussed in greater detail and depth in Chapter Five which features discussion of findings and the wider meaning and implications of these findings.

## Analysis of data derived from Method 1: the adaptive comparative judgement workshop and subsequent individual judging:

| Teacher | Infit | Local | Median Time |
|---|---|---|---|
| Teacher 1 | 0.72 | 110 | 32.2s |
| Teacher 2 | 1.18 | 12 | 122.1s |
| Teacher 3 | 0.8 | 110 | 3.3s |
| Teacher 4 | 0.66 | 110 | 149.6s |
| Teacher 5 | 1 | 24 | 175.8s |
| Teacher 6 | 1.38 | 110 | 3.5s |
| Teacher 7 | 0.88 | 70 | 3.3s |

*Method 1 Comparative Assessment Judgement trial judge results*

**Infit:** This metric represents the level of agreement between judges on scripts, with respect to the overall quality of the scripts that teachers were judging.

We can note from infit measure in this method that the judging decisions varied significantly between different judges. Of the seven judges that participated, five were deemed by the NoMoreMarking software to reaching judgement decisions that were consistent with that of their peers, owing to their infit score being at 1.0 or lower. The two remaining teachers scored over 1.0, with infit ratings of 1.18 (teacher 2) and 1.38 (teacher 6) respectively. Teacher 6 is of particular interest here, in that their infit score falls into the 'inconsistent' category. A high infit score, as evident with this teacher, is not necessarily an indication of negligence or poor performance in the judging of script quality; it could indeed indicate the opposite, in that this teacher is a highly competent judge of script quality, and the remaining judges are less competent in comparison, hence the high inconsistency score. What we can recognise the score as providing is a spotlight onto the important theme of agreement and standardisation of assessment decisions. This will be explored further in Chapter Five.

**Local:** This metric represents how many judgements each teacher made in one hour.

The total number of judgements recommended by NoMoreMarking in view of the sample size was 110. There was a variance in the number of judgements the teachers in the sample completed: four teachers made 110 judgements, one made

70 and the other two completed 12 and 24 judgements respectively. The mean average across the entire sample was 78 judgements across the seven teachers. It is important to note that the teachers that took part in this trial were not incentivised to take part and were not asked to undertake this as a requirement of their role at the college. Ultimately, those that participated did so of their own choosing and elected to take on this additional task. These scores need to be read with an appreciation of these contextual factors.

In the workshop sessions in which this judging sample was launched, no teachers fulfilled their allocation of judgements. What this result shows is that the four teachers who completed all 110 judgements assigned to them did so by undertaking individual judging at a later date. By the same token we can recognise that the three teachers that did not complete their allocated number of judgements did not undertake any additional judging beyond the workshop session. This might have been for a number of reasons, including a lack of time, not seeing the benefit, or simply forgetting to do so. Enquiring about teacher impressions of using this approach to assessment is important if we are to better understand the reasons and motivations for the number of local judgements completed in this trial, and the perceived usefulness and value of undertaking ACJ. This is a theme that is explored and discussed below in method 3, in the semi-structured interviews with teachers.

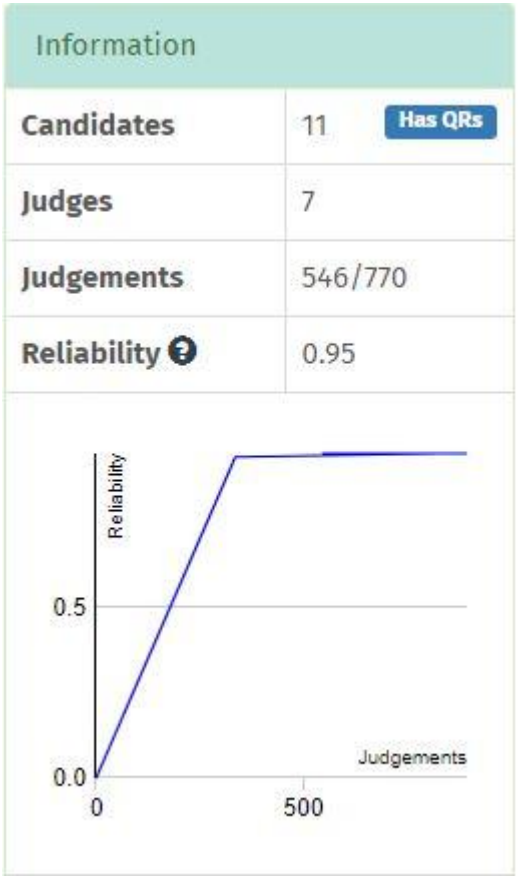
**Median time:** This metric shows the duration of time spent judging each script individually.

There is significant variance evident here between teachers, with the shortest

median time at 3.3 seconds and the longest at 175.8 seconds. We can observe that teacher 2 and teacher 5 have proportionally higher median times than the large majority of teachers owing to the fact they have completed far less judgements, and as such were likely still familiarising themselves with the scripts when early into their judgement sample. This indicates that, on average, the teachers in this trial gained in speed when forming judgements regarding the quality of the scripts they were reading. We can note that four out of five teachers that completed at least 70 judgements had a median time of 33 seconds or less, with three of those having a median time of 3.5 seconds or under.

The significant outlier here is the median result of teacher 4, who completed 110 judgements with a median time of just under 150 seconds. The result here indicates that teacher 4 took far longer in forming their judgements per script pairing than when compared with their peers, and that a duration in excess of 150 seconds, or 2 ½ minutes was taken for at least fifty-four of the one-hundred and ten judgement decisions that they made. This is a remarkable disparity when compared with the rest of the judges' median results and raises questions about the process of assessment they undertook. For example, was this judge analysing each script in more detail than the other judges? Where they re-reading each in script in depth in each assessment iteration? A key question that arises here in view of this result is does the average time spent judging texts have an impact on the quality of the judgement, that is to say, the judges infit score?

**Reliability**




*Method 1 Comparative Assessment Judgement trial results overview*

Reliability of the assessment decisions in the trial scored very high, with a reliability rating of 0.95. This was despite the total number of judgements reaching 546, short of the 770 total that was recommended. Evident here is how the reliability of assessment decisions, and subsequent scaling of scripts in order of quality, is calculated with respect to all decisions that are made by the judges. The NoMoreMarking software contains within its sorting algorithms a standardisation of judgements through which outlier judges and judgement decisions are identified and considered within the sample, but are not given the licence to affect the overall reliability rating. This is not to say that outliers and anomalous judgements are

desirable for the sample, but rather that the accuracy of the ordering of scripts in terms of quality is reflective of the majority consensus in any given sample.

## Analysis of data derived from method 2: the adaptive comparative judgement workshop:

| Teacher | Infit | Local | Median Time |
|---|---|---|---|
| Teacher 8 | 0.91 | 71 | 26.0s |
| Teacher 9 | 0.72 | 96 | 16.5s |
| Teacher 10 | 1.03 | 80 | 18.1s |
| Teacher 11 | 0.79 | 71 | 11.6s |
| Teacher 12 | 1.47 | 70 | 14.8s |

*Method 2 Comparative Assessment Judgement workshop judge results*

**Infit:**

We can note from this sample that the majority of judges were largely in agreement with one another regarding their judgement decisions, with teacher 12 the only teacher in the sample to show a significant difference to others.

**Local:**

All teachers achieved at least seventy judgements (the number recommended for the sample size), and the mean average was 77.6 judgements across the five teachers. Significantly, all teachers completed at least 70 judgements in the one hour allocated to them during the workshop.

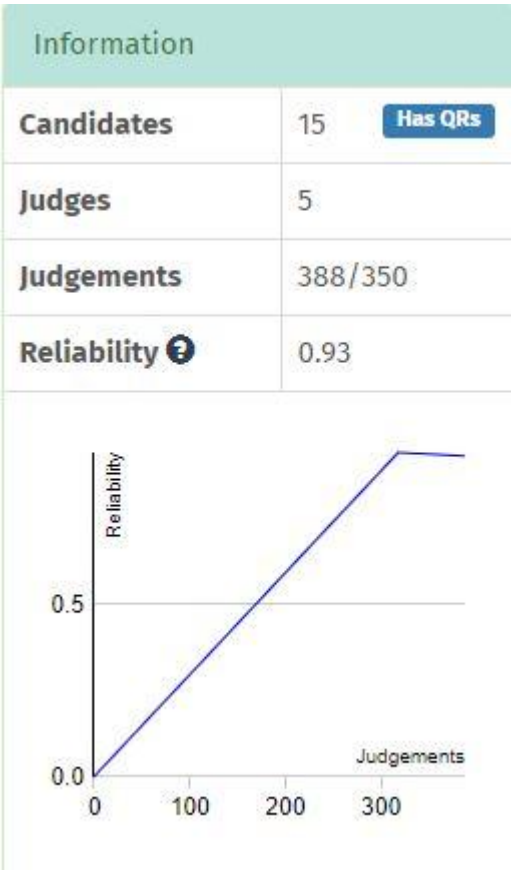
**Median time:**

There is some variance evident here between teachers, with the shortest median time at 11.6 seconds and the longest at 26 seconds. There is also far less disparity between the median judging times when compared with method 1; while the median times of method 1 had a range of 172.5 seconds, the range of median times in method 2 was 14.4 seconds.

One possible contributing factor to the conformity in median judgement times was the environment in which the comparative judgement was taking place. Whereas in method 1 teachers undertook a percentage of their judgement decisions in a workshop setting before being given the opportunity to complete their sample in an environment of their choosing and in their own time, the teachers in method 2 shared an environment and given a duration in which to complete as many judgements as they could. It is possible that by sharing an environment the teachers that participated in method 2 aligned their assessment practices to one another, in respect of how long they were taking to make their judgements. We can observe that the teachers in method 2 were undertaking assessments of creative writing using an

unfamiliar online system without the aid of a mark scheme or assessment standards, and despite them working individually, there might have been a desire, consciously or not, to conform to the standards that were being dictated by the group. There might, for example, have been a teacher that would have felt more comfortable taking less or more time in making judgements but did not want to be perceived as being too hasty or ponderous with their judgements. This raises questions about the role and importance of environment when undertaking ACJ activities as a group.

**Reliability**




*Method 2 Comparative Assessment Judgement trial results overview*

Reliability of the assessment decisions in the trial scored very high, with a reliability rating of 0.93. A similarity between the reliability result evident in method 1 and this

method can be seen, with their overall reliability being largely unaffected by one judge's high inconsistency in their infit score.

# Chapter 4: Emerging Themes and Findings

## Methods 1 & 2: questions to consider and emerging themes

The findings from the trials in methods 1 & 2 presented in Chapter Three uncover a number of emerging key themes and raise questions that are worth exploring in greater depth and detail as we seek to understand more comprehensively the challenges and benefits of using ACJ to assess the quality of student creative writing.

### What makes good creative writing?

Firstly, we can observe that the infit score provided by the NoMoreMarking software provides an account of teacher agreement of script quality, and by extension an indication of how standardised teachers' judgements are across a sample. The large majority of teachers having a consistent infit score, as seen in the methods above, indicate that these individuals share a consistent construct of what makes a good quality item of creative writing. As both of these ACJ activities were completed without the aid of mark schemes or assessment standards it is significant that the judging collective in both methods achieved a high reliability rating as to the quality of the creative writing scripts. This indicates that in both methods judges had a shared construct of what makes good creative writing, with sufficient confidence to

allow them to consistently identify the better of two texts when presented with the choice. With this in mind, we can ask questions to advance this line of enquiry:

- What makes good creative writing at GCSE English level? What did the teachers in these methods identify as contributing towards a proficient creative writing script?

With largely consistent infit scores and a high reliability score evident in both methods we can explore this question more acutely by asking further questions. Firstly:

- What commonalities were identified across multiple teachers? Is there a shared construct of what makes proficient creative writing?

We can note that judges in both methods had an outlier infit score, and these too are worth further consideration, as from these we can take that their construct of good quality might differ from that of their peers. From this we can pose:

- What unique aspects were identified by individual teachers?

Method 3, the semi-structured interviews with teachers, will explore the theme of what makes good creative writing. The findings gathered through these will be considered alongside these questions in Chapter Five.

**Judging consistency compared with experience**

Through method 2 and the semi-structured interviews with teachers it is possible to build a profile of each teacher's experience in teaching GCSE English in the FE sector. This includes gaining information about their continuing professional development as a subject specialist teacher. From this it is possible to examine the relationship between teacher experience and judgement consistency in respect of overall sample. The question to consider here is:

- Is there a correlation between how experienced a teacher is in teaching GCSE English and the consistency of their judgements compared with their peers?

**Judging consistency compared with duration per judgement**

Following from the above infit score, we can note that there is some evidence from methods 1 and 2 to suggest there is a correlation between time spent and the agreement between judgements in consideration of other judges' decisions. The most accurate judge across both trials, teacher 4 in method 1, had the longest median time of all judges across both methods. This is perhaps to be expected. But to correlate a longer time spent judging with a greater reliability of judgement is too simplistic. It is evident from other individual teacher results that a consistent infit score can be attained with a short median judging time. This invites us to consider the relationship between judgement reliability compared with duration per judgement. We can enquire:

- What is the relationship between judgement reliability compared with duration per judgement? What factors are important to consider in this relationship? How might it differ between different judges?

Moreover, there are pragmatic considerations to make as well. It is perhaps to be expected that more time spent making judgements leads to greater consistency. In practice, however, teachers do not have an abundance of time for the assessment of student work. Accordingly, we can enquire:

- Is there an optimal amount of time that judges should be spending on average when judging creative writing scripts of this type? What compensating factors need to be considered in reporting an optimal time?

These themes and questions will be revisited and examined in more detail in Chapter Five.

# Findings from method 3 - Semi-structured interviews with teachers

The findings from the semi-structured interviews with teachers are reported below in three sections: (1) experience and training in teaching and assessing GCSE English, (2) reflecting on the use of adaptive comparative judgement, and (3) the practice of undertaking comparative judgement.

## (1) Teacher experience and training in teaching and assessing GCSE English

The table below features a summary of the findings from interviews with teachers regarding their experiences of teaching and assessing GCSE English, and any formal or informal relevant training they have undertaken alongside this. The information shared by teachers and presented here is self-reported, and this should be considered when forming any conclusions. Also included in the right-most column in the table is each teacher's infit score attained through the ACJ workshop in method 2, to allow for comparisons to be drawn between teacher experience and training and their consistency in judging in view of their peers.

| Teacher | Experience in teaching GCSE English | Formal training | Informal training | Infit score |
|---|---|---|---|---|
| **Teacher 8** | 4 years in Further Education;<br><br>10 years preceding this in secondary school settings as an ad hoc GCSE English cover supervisor | Attended an AQA event when new specification was introduced (2016);<br><br>English subject specialism one day training event - "stepping stone" to a Level 5 subject specialist qualification | Participating in, and leading, standardisation meetings for the college's GCSE English team (multiple years);<br><br>Working individually with teachers to standardise and IV marking (one academic year) | 0.91 |
| **Teacher 9** | 3 years in Further Education;<br><br>Some experience of teaching GCSE English (old specification) with the YMCA prior to working in FE | Attended an AQA event when new specification was introduced (2016)<br><br>English subject specialism one day training event - "stepping stone" to a Level 5 subject specialist qualification | Participating in standardisation meetings for the college's GCSE English team (multiple years); | 0.72 |
| **Teacher 10** | 4 years in Further Education, comprising 3 years in an adult college (only GCSE English teacher there) and 1 year in current institution; | None. Attended some AQA briefings on qualification changes, but nothing practice focused. | Informal discussions with other teachers (self-directed) | 1.03 |
| **Teacher 11** | 2 years in Further Education | Attended a 1-day AQA event focused on planning and delivery of GCSE English | Elements taken from in-house staff development days (e.g. questioning techniques, and how these can be applied to GCSE English teaching) | 0.79 |
| **Teacher 12** | 2 years as a trainee teacher in Further Education (teaching 2.5 hours per week supervised by a mentor) | PGCE course (no specific elements cited)<br><br>Participating in standardisation meetings for the college's GCSE English team (one event, four hours in duration) | Conversations with PGCE mentor | 1.47 |

Examining each teacher's reported experiences and training profile in GCSE English and comparing these with the infit scores attained through method 2 reveals that there is a correlation between the two. We can observe that the least experienced member of the teaching team that contributed to the ACJ workshop, Teacher 12, had the least agreement with their peers' judgements. The other four teachers had relatively consistent decisions in comparison.

On formal training, teachers shared similar experiences. Many had participated in training led by the GCSE English awarding body provider, AQA. Teachers 8 and 9 had participated in a Level 5 subject specialism stepping stone training event that was delivered in the college. There were interesting reflections shared when teachers were asked to report on the effectiveness of the formal training they had participated in as GCSE English teachers, represented below:

***Teacher 9:*** *it* (the AQA training) *wasn't enough really because...it's one of those things...until you've done a whole cycle of it you don't really know how you're doing. What would have been really useful is if we'd marked the mocks and then AQA had marked the mocks...and then they could have come back higher, or said you need more of this or more of that.*

***Interviewer:*** *...and what would that have given you, that mock marking?*

***Teacher 9:*** *that would have given me an insight if I was going in the right direction or not. Not necessarily if the student is going to pass or not, but if I'm doing the right things.*

---

***Teacher 10:*** *what was more beneficial to me was talking to other teachers...we did share things on the creative writing front.*

---

***Teacher 11:*** *I found parts of it* (the AQA training) *beneficial...the downside of it was that it was catered to GCSE English teachers in secondary school where the idea is you can start teaching elements of the topic from as early as year 7 or 8, and it was presented as everything is a slow burn...whereas here we have to do it in 30 weeks at the college...so there were elements that didn't really work for FE...it was very good being able to talk to other practitioners and share ideas, things like that.*

---

***Teacher 12:*** *for the staff development days and that sort of thing* (referring to a GCSE English team standardisation meeting) *I think it was interesting to see how other people do it, I got to see the feedback sheets, their marking...but I also got to see the standard of marking quality that the college asks for.*

Many of the teachers shared that they felt that the formal training they had participated in was not as effective as they had hoped it would be, or felt it could have been improved in some way. This was a result of the training lacking a focus on individual assessment practice (teacher 9), an overemphasis on operational rather than pedagogical matters in training led by the qualification awarding body (teacher 10), or that training that was more tailored to teaching GCSE English in secondary school settings (teacher 11).

The large majority of teachers commonly identified dialogue and collaborating with colleagues at these formal training events as a particularly beneficial element. Few

references, if any, were made to the content of the training event itself. Further to this, there was evidence provided by Teacher 12 that indicated collaborating with colleagues provided insight for them into the ' standard of marking quality the college asks for'. This points to the craft-like nature of learning aspects of a teacher's role, in this case the decoding of assessment standards, by looking to more experienced colleagues to provide guidance and model effective practice.

On informal training, teachers identified standardisation events and discussions with colleagues as the most common forms they had participated in. On how effective teachers felt these forms of training had been, contributions included:

***Teacher 9:*** *we've done standardisation meetings amongst ourselves. A lot of disagreement on those...it might have been because there was a couple of strong personalities there...but again, because all had differences of opinion and different ways of doing it, so I'm not sure how useful it was. It made you think about your own practice because you heard what other people said, but you still didn't necessarily agree with them. That's why it would have been nice to have an overarching AQA (sic) saying "this is what we want".*

---

***Teacher 10:*** *I think here you might only have a chat with another teacher...and it's literally just a chat with another teacher. What was interesting for me was there was a new teacher that started this year and we were doing a different training and we got chatting, at the beginning of the academic year...and she said she felt as though she was on her own...and I could relate to that as when I started here I was just left to get on with it, and you kind of are out on a limb a bit. And what she'd found really*

*odd is that when working at a different college everyone sat in a room and did the scheme of work so it was uniformly delivered…but we don't do that here…*

---

***Interviewer:*** *so how effective were these conversations, are these conversations, in equipping you with…?*

***Teacher 12:*** *oh they were vital. They gave me a lot of grounded context. Like, the formal educational stuff that came from the PGCE was obviously very useful, and I wouldn't have been able to advance without it, but talking to my mentors...that was how I really learnt where it really applied.*

---

Teachers shared a range of reflections on informal training. Teacher 9 highlighted the difficulties they and the team often encounter when trying to agree a standard with just one unified voice, even going as far as to state that AQA should provide clear guidance on what they want from marking uniformity. As already presented in Chapters One & Two, the subjectivities inherent in the AQA marking schemes and assessment standards for creative writing make this a difficult feat indeed. Nonetheless, Teacher 9's contribution exemplifies the perceived lack of resolution that is sometimes evident in standardisation activities that they are participating in. Teacher 10 highlighted discussions as an important training activity in the induction of new teachers to the college. The example was shared of a new teacher joining the college having taught at another institution elsewhere beforehand, and how they needed guidance in order to introduce them to the working practices that were different to their previous context. In a similar vein, Teacher 12 spoke of learning from mentors and how this helped them gain a "grounded context" where knowledge

and understanding could be applied. These two contributions again point to the craft-like nature of professional learning, and how more experienced colleagues play a crucial, if not formalised, role in establishing and maintaining of standards.

## (2) Reflecting on the use of adaptive comparative judgement

Excerpts from each individual judge's semi-structured interview are reported individually in the following section, and are each concluded with a summarising commentary that draws together key themes.

**Teacher 8:**

***Teacher 8: -*** *"What I was looking for initially at was being engaged and being interested in the structure and what was going on [...] the content. And then looking at the sentence structure and SPAG (spelling, punctuation and grammar) after."*

---

***Teacher 8: -*** *"With some of them (the texts) it wasn't clear what was going on because the sentence structure was so bad, so it was a bit of both really. But if it was semi-decent I was judging it on how engaged I was first"*

***Interviewer: -*** *"and that's the measure for you that matters?"*

***Teacher 8:*** *"Yes. If I can read something from the first word to the last work without going back over a paragraph to figure out what's going on that's a good piece of writing to me...it's not gone off on a tangent, I've not thought "I don't know where you're going with this…"*

---

***Teacher 8: -*** *"I struggle with the mark scheme sometimes. Just because they haven't put a semi-colon in there, just because their language is more simplistic and not sophisticated, doesn't mean it isn't a really good piece of writing"*

***Interviewer:*** *"Do you value that [text] cohesion more than other aspects?"*

***Teacher 8:*** *"Yeh, I think I probably do, because our students struggle to have ideas and struggle to be creative, so if they've created something that is cohesive and interesting…we're talking about FE here, and students that are vocabulary poor...how can we be expecting students to use that if they've managed to write something from start to finish that's engaging and fit for purpose? For me I'd want to give them a pass straight away but we've got to stick to the mark scheme, which is unfortunate."*

---

**Teacher 8 - summarising commentary**

Evident in these reflections was how teacher 8 adopted the perspective of a reader more so than one of an assessor, noting that they were looking for *"engagement*" in the text, and how the content created and sustained *"interest"*. More technical elements such as spelling and punctuation were secondary considerations on secondary reading. Teacher 8 also spoke about how mark schemes use limiting elements that require students to use specific technical elements in their creative writing, such as *"sophisticated language",* in order to be deemed at a good standard. They reflected on their *"struggle"* to mark in this manner, explaining that they value textual cohesion that makes for interesting writing.

Interestingly, Teacher 8 spoke of the difficulty of using mark schemes to assess creative writing, and found adopting a personal interpretive perspective to be more effective as a way of determining textual quality. While it might be effective, as was seemingly so in this case, we can note how a personal interpretive approach to assessment in this manner might suffer from a lack of transparency, reliability and consistency. A judge asking themselves whether a creative writing text is interesting is a valid question to pose, but cannot be relied on as the sole indicator of quality owing to the seemingly vast disparity in what different judges would find and agree to be interesting. The idea of textual interest remains a valid consideration in the context of this enquiry on account of the agreement seen in judges when determining script quality, and will be explored further in Chapter Five.

**Teacher 9:**

***Teacher 9****: - "It would be great to have a benchmarking activity at the start of the year - what is a (grade) '3', a '4', a '5'? This would help us and the students."*

---

***Teacher 9****: - "It's a lot better than sitting there with a mark scheme, which can drive you up the wall sometimes because you sort of know where to put a piece when you look at it, and then see how the mark scheme fits around it."*

---

***Teacher 9****: - "It changes what I was looking at. There might be a few spelling mistakes but the actual content is really good, and I really think that the mark scheme - you don't always look at the content...the other bits and pieces...you've got to tick the boxes - whereas when you read it cold you think was that a good story?*

*did it grip me? Was it well-written? Did it flow? And that's all you concentrate on really...and that's all you're looking at."*

---

**Teacher 9 - summarising commentary**

Teacher 9 reflected on the practical application of ACJ, and how it might be used to enable standardisation and benchmarking of specific texts appropriate to specific levels. This could be completed by the teaching team and then be shared with students to provide them with models pertaining to different levels of performance. An interesting relationship between scripts and mark schemes was discussed, with teacher 9 noting *"you sort of know where to put a piece (in reference to the level it would be awarded) when you look at it, and then see how the mark scheme fits around it"*. It appears from this that this teacher has previously adopted a similar approach to assessment as has been encouraged in this enquiry through adaptive comparative writing, in which a tacit understanding of good work contributes to the judgement, alongside or in favour of a mark schemes codified standards.
Also apparent in these reflections was the use of figurative language and metaphor to articulate the intangible qualities they valued in creative writing texts, as seen in the examples 'did it *grip* me?' and 'did it *flow*?' These findings give insight into how this teacher articulates their tacit understanding of what good creative writing is beyond the specific criteria featured in assessment standards.

**Teacher 10:**

***Teacher 10****: - "In English we're assessing against a mark scheme against all the criteria, SPAG and all the rest of it, but actually sometimes when you're creative that kind of goes out the window, because you're not thinking in a uniform way. So a silly example is starting a piece of creative writing with 'but' or 'and', in the context of a piece of creative writing it works."*

---

***Teacher 10:*** *"I think you have to divide yourself - are you looking at it purely in terms of creativity? Or are you looking at it in terms of good English?"*

***Interviewer:*** *"which do you value more?"*

***Teacher 10:*** *"creativity"*

***Interviewer:*** *"why is that?"*

***Teacher 10:*** *"because it's more interesting."*

***Interviewer:*** *"but is that what matters for learners?"*

***Teacher 10:*** *"I think if you can get them to use their imagination and start to tap into that resource, that pays bigger dividends for them in the long run because they're engaged...if you're going to keep going on about SPAG...and don't get me wrong, that's important...but if you're getting them to unlock something then I think that can come later...we can tidy up later (on technical accuracy in writing)...but the creativity stuff, you need to get them not to be scared of it and accept that they've got it. Some of them say…"I can't do creative writing Miss". - "Well, yes you can" - we just need to find a way to unlock it to help them express themselves, and tidy up after."*

***Interviewer:*** *"and when you say unlock 'it' - what is 'it'?"*

***Teacher 10:*** *"creativity, potential"*

***Interviewer:*** *"and what is that?"*

***Teacher 10:*** *"freedom to write whatever you like and not be worrying what people think about it*."

---

**Teacher 10 - summarising commentary**

Teacher 10 spoke of how good creative writing can often subvert normal conventions in grammar and structure, citing the example *"starting a piece of creative writing with 'but' or 'and'"*. The example was discussed in reference to mark schemes and their focusing on specific technical elements of grammar, and by extension to very conventions of language that teachers of GCSE English teach to their students, noting that *"when you're creative that kind of goes out the window".*

Creativity was a recurring theme in teacher 10's interview. They spoke at length of the idea of creativity as a *"resource*" that could be *"unlocked"* in students, stating that this was far more important to foster than technical accuracy which could come later once enthusiasm for the subject had been developed. Whilst somewhat tangential to the practice of assessment through adaptive comparative judgement, the idea of creativity as a resource and student engagement in their studies is relevant to this enquiry. The teacher's concluding point stated a goal of teaching GCSE English should be to enable students the *"freedom to write whatever you like and not be worrying what people think about it".* There are significant pedagogical

considerations that we need to appreciate that follow from this. In a system in which student performance in any given task, including creative writing, is measured and assigned a numerical value to account for its quality, it is perhaps difficult to see how students could simply remove themselves from *"worrying what people think about it".* Whether student concerns would be as pronounced in an assessment environment in which comparative judgement was used, where judgements are made against other texts and not through an external standard, is perhaps worthy further consideration.

---

**Teacher 11:**

***Teacher 11:*** *"It allowed you to take a moment and appreciate it as a piece of creative writing, rather than immediately going in for the critiquing of everything from the mark scheme...making sure students had ticked all the boxes."*

---

***Teacher 11:*** *"I think students would like the feedback that teachers enjoyed what they wrote, rather than what you'd got i.e. "you got your grade 3" or things like that."*

---

***Teacher 11:*** *"With a mark scheme you break it down at an earlier stage, it seems as though you're compartmentalising it in a way, and you're making notes along the way to see if they're getting marks or missing marks. The comparative judgement gives you that opportunity at the beginning to take it all in as a whole, because it's asking "which one is better?", and that's far easier than having to tear it down to its constituent parts"*

**Teacher 11 - summarising commentary**

On using ACJ teacher 11 noted how it enabled them to view texts as a whole when considering their quality, rather than checking if individual itemised elements from the mark scheme had been fulfilled. This extended to being able to consider if they enjoyed the text, a consideration in judging quality that was also mentioned by teacher 8 in their interview. Teacher 11 extended this idea further in noting that they felt students might like to receive feedback from teachers stating things as simply as they *"enjoyed what students wrote*". This chimes with D'Arcy's (1999) concept of dialogic feedback discussed in Chapter Two. Chapter Five will build on these possibilities further, by exploring in more detail the shift towards whole text appreciation that ACJ allows for.

**Teacher 12**

***Teacher 12:*** *"I think it's a great method, particularly once you're at the level of a professional teacher, or experienced teacher, where you've got knowledge of what makes a good answer [...] and that's at more of an instinctual level where you wouldn't need to check back against a mark sheet or comb through it for every little detail. You just know whether it's a good answer or a bad answer [when comparing with other scripts]."*

---

***Teacher 12:*** *"It's as much a matter of feeling. I don't think Hemingway would pass most creative writing courses because he's too short spoken, but we agree that he's someone of quality writing."*

---

***Teacher 12:*** *"This option represents it more in the way that it feels to a reader, which is what really matters when you're writing creatively...how it comes across to people reading it. It's not about expressing facts...it's about expressing a feeling or working a theme or idea."*

---

**Teacher 12 - summarising commentary**

Teacher 12 felt that there was a need to attain the *"knowledge of what makes a good answer...an instinctual level*" in order to conduct ACJ effectively, pointing to the importance of first developing, and then drawing upon, a tacit understanding of good quality creative writing in which *"you wouldn't need to check against a mark scheme."* This idea was extended with *"it's as much a matter of feeling"*, citing instinct and other intangible indicators as being crucial in this process. The example was given of Ernest Hemingway, whose writing is much celebrated but, it was argued, would fall short of meeting the success criteria laid out by some creative writing standards. This example demonstrates effectively a point also made by teacher 10 on how good creative writing can often subvert convention and flout existing standards to its own benefit, by doing so creating an identity that defies cliché or prosaicness. What we can resolve from this is that adaptive comparative judgement is an effective assessment approach in enabling teachers to determine textual quality, largely owing to the crucial role that tacit knowledge plays in forming such judgements.

## (3) The practice of undertaking comparative judgement

Two questions were asked of teachers that focused on the practice of undertaking comparative judgement. Excerpts from the responses to these are reported below:

**What helped you arrive at the decision?**

***Teacher 8: (****referring to one of the texts in front of them) I enjoyed it more. It has suspense, it's structurally much more engaging than text B. We've got a character in here, we've got interest...it's also fairly well structured sentence and punctuation wise. Grammatically it's quite sound too.*

---

***Teacher 9:*** *"It's a mixture of the flow and the content, really. As an English teacher when there's glaring errors they leap out at you sometimes and it sort of interrupts the flow of your thoughts...I'm not looking for if somebody has spelt something wrong, it doesn't matter, but it's flow I'm looking for as much as anything, and unusual images and not just normal sorts of word patterns, that sort of thing."*

***Interviewer:*** *"Could you break down this idea of 'flow'?"*

***Teacher 9:*** *"it's a thing that...I don't know...if something jars with the rest, and that's it really. It's a difficult one to quantify really. I think it's your instinct really, and what you like reading.*

---

**What is it you're drawing on?**

***Teacher 8:*** *(laughing)....a feeling. It's like reading anything. Some things are interesting to read and some things are difficult.*

***Teacher 10:*** *Probably experience. I've read a lot of books and if a book doesn't interest me...doesn't pull me in...then I'm not interested. When you read something there has to be a draw...there has to be something to pull you in. If you're reading something and it's making you think - "why is that happening" - "why is she doing that?" - it compels you to read on. Some of it is does it interest me...does it connect with me? ...and text A definitely does more than B.*

---

***Teacher 11:*** *There was still an element where I was thinking of the mark scheme in the back of my mind, in relation to spelling and grammar. But in terms of the content I was reading through and thinking which one did I enjoy most, what one is the most complete story, that held my attention more and made me want to read on to the end, which is something that we should be encouraging more in our learners.*

---

**Commentary on the responses to "what helped you arrive at the decision?"**

Several references were made by teachers to intangible qualities, including "*suspense", "structural engagement", "flow" and "content*". Each of the elements cited here describe things that appear throughout a script, rather than in isolation. Moreover, they are achieved through the successful marriage of a combination of techniques and structural decisions that together contribute to a greater whole. If we are to take these elements as being indicative of good creative writing, and accordingly as those that students should be shown and learn to apply in their own writing, it follows that there are significant implications for the teaching of creative writing in GCSE English settings. In this respect, teachers must make conscious efforts to expose students to models of creative writing that achieve these effects

successfully, and to explicitly signpost where and how these effects are being achieved.

Textual "*flow"* is an interesting example that perfectly demonstrates this. As previously discussed above, *flow* exploits a metaphorical construction to account for how a text feels when it is being read. If a text *flows* well, we can understand it to be easier to read and follow the meaning of; conversely if it does not flow we can take that the opposite is true, and that there are perhaps awkward word choices or out-of-place structural devices that detract from its fluency. But the problem arises if we trace the root meaning back to the choice of a metaphor to account for this. Metaphor is used to describe something through comparison with another object, action or phenomenon to which it is not literally applicable. Accordingly, we can recognise that a metaphor has been used here because of the difficulty in articulating what *flow* actually is, hence the need for a metaphor to provide a frame of reference that others can associate meaning to. The challenge in view of this that teachers of GCSE English must navigate is how these elements, such as *flow*, can be taught to students. These challenges, and the tendency for metaphor to be used when accounting for different elements of textual quality will be explored in greater depth in Chapter Six.

**Commentary on the responses to "what is it you're drawing on?"**

The responses to this question reveal the extent to which teachers adopt the position of a reader, in addition to that of an assessor, when forming a judgement when using comparative judgement. All teachers stated that they felt that 'engaging' with a text

was an important indicator of quality and was what they valued over other technical aspects. The student scripts were appreciated in a holistic way, as authentic artefacts rather than items created for the sole purpose of measuring performance in creative writing. In this sense, there was evidence of teachers adopting a more dialogic approach to textual engagement when using comparative judgement, as seen in questions such as “why is that happening?” This example, and others like it, appear to indicate at how ACJ might be used as a vehicle to facilitate assessment *for* learning, in which teachers and learners exchange dialogue about the creative decisions made and reasons for them. Again, this resonates with the work of D’Arcy (1999) who advocates the use of ‘interpretive responses’ to creative writing, in which a reader ‘adopts a meaning-related paradigm would be prepared to take an aesthetic stance to the text, prepared to engage with it, imaginatively, empathetically, and visually’ (1999:14).

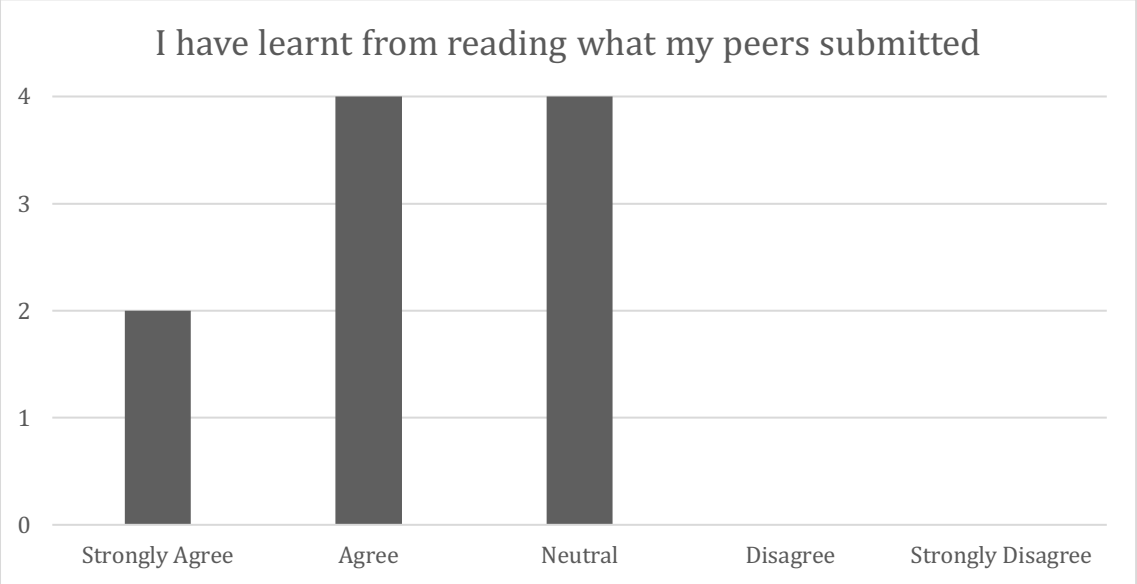
## Analysis of data derived from method 4: student questionnaire

The final two methods reported below account for the activities that sought to gain insight into student perceptions of using adaptive comparative judgement as an approach to peer assessing creative writing scripts.

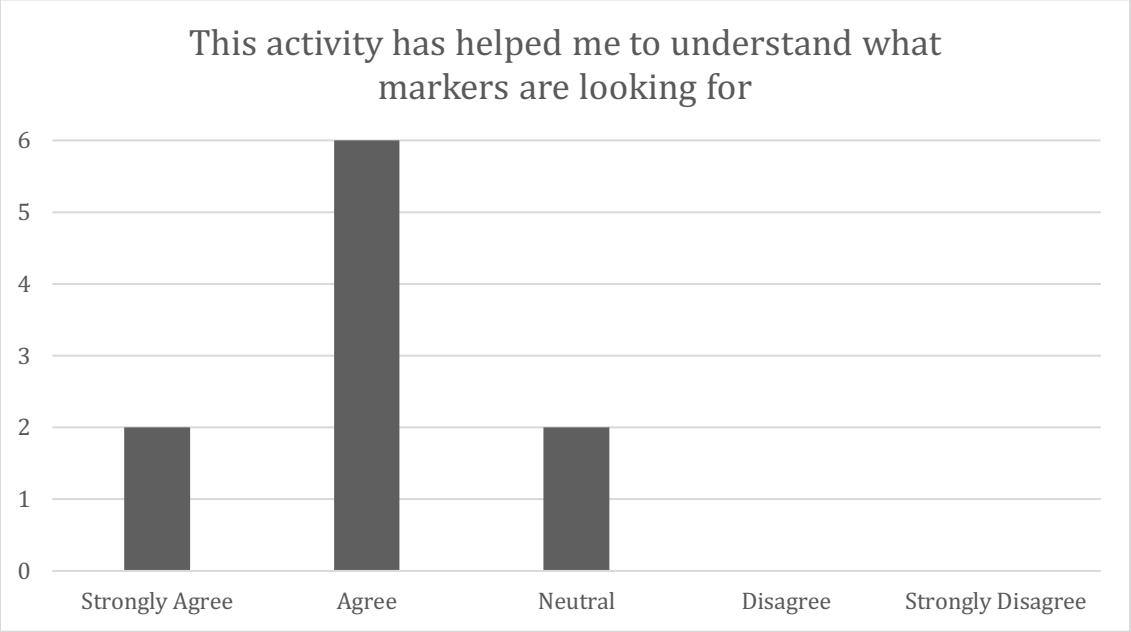
Ten students completed this questionnaire after completing a short ACJ activity. Findings from method 4 are split into four sections: the perceived value of ACJ as a method of peer learning (1), ACJ as helping to develop an understanding of the subject (2), and if it was an effective use of time (3), and a free comment section (4).

These sections are followed by a commentary summarising the key themes emerging.

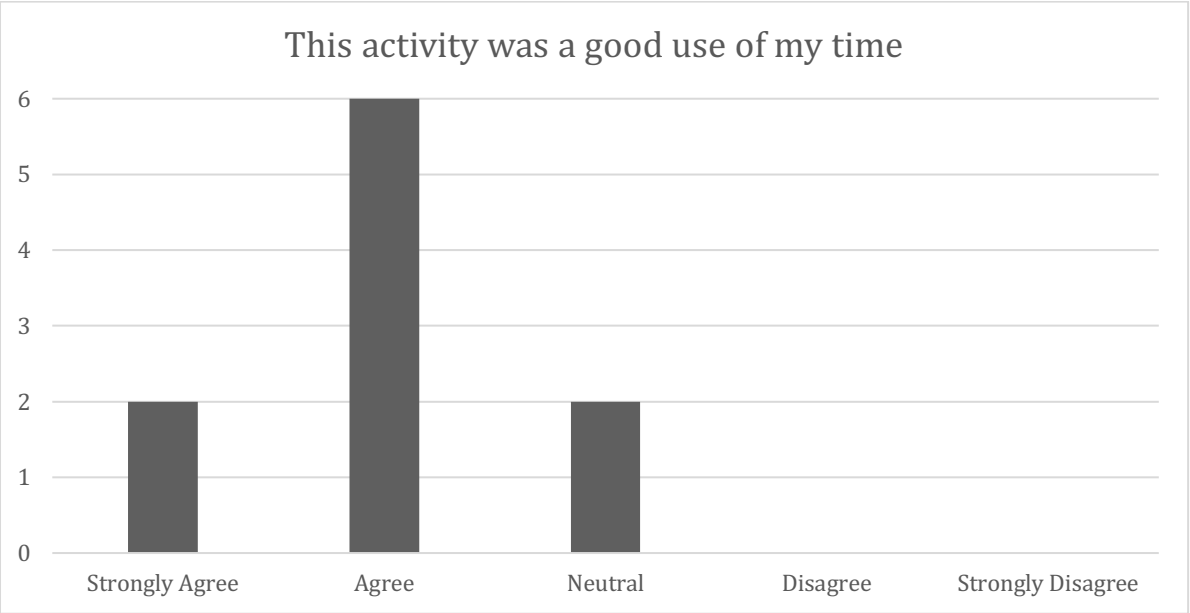
## The perceived value of ACJ as a method of peer learning (1)




## ACJ as helping to develop an understanding of the subject (2)



## If it was an effective use of time (3)




## Feedback from the free comment section (4)

*“Gave us an opportunity to take a look at others work”*

*“Some text I found really useful where they start the sentence with adverb.”*

*“It was good to look at a lot of work done by students and see how, what and why they wrote this. I also learnt what level I am compared to many of the student are* [sic]*. I am at a good level.”*

*“It was an eye-opening experience. some I couldn't read others I wish I hadn't”*

*“this activity has helped me to develop my writing skill by reading different pieces of extract.”*

*“I have learnt there is a great different level of work from each student. Also how creative some students are with there* [sic] *work.”*

*“If the answer was written on the computer, would have been better to understand. It was good to see other answers and then compare to your own one to see if you can make any improvement.”*

*“I got some ideas how to be creative in writing and for what to looking for* [sic] *when you have to judge the text”*

*“It is a good exercise to understand other people's style of writing. Very useful.”*

## Emerging findings from method 4

Six of the ten students responded positively when asked if they had learnt from reading what their peers had submitted. The remaining four responded with a ‘neutral’, suggesting they neither agreed nor disagreed. This indicates that a majority of students perceived ACJ to result in them learning from reading what their peers had written. These views were corroborated through the free ranging comments shared by students at the end of the questionnaire, which are discussed in more detail below. But these views were not uniformly shared across all ten students. We can trace several possible reasons for why some students felt they did not learn from reading their peers’ work: students might have lacked confidence when considering which script of the two they felt was better owing to the omission of standards or a mark scheme to guide them; they might have struggled to read the scripts they were judging, or felt that they offered little to them in terms of modelling of good creative writing; or they might have not have sufficient time or inclination to reflect on the value of the task despite any potential merits it had offered, and felt that they did not learn anything from the process as a result.

An emerging theme here is the notion of students assuming the identity of an assessor, distinct from that of a student, in which they are given licence to form judgements on the quality of their peer's work. In this tradition the judgements that students arrive at are regarded as valid and not supplanted by that of a teacher or expert, because the student is regarded as a competent judge of quality in their own right. Accordingly, students are empowered to assume this identity, and recognise themselves the validity that their judgements carry. But the conditions in which peer assessment of this kind can take place are not realised without preparatory work. Marshall and Wiliam (2006) observe that 'the teacher has to create a safe environment in which pupils feel comfortable having others read their work, collaboration and share of practice have become the norm...and pupils need to see examples of good practice to be able to know what to do" (2006:19). This idea of students as peer assessors is fundamental to this discussion on adaptive comparative judgement as a way of understanding what students think good creative writing comprises, and will be revisited and explored in greater depth in Chapter Five.

Eight of the ten students responding to the questionnaire found the process of undertaking comparative judgement to help them understand what markers were looking for in creative writing (2), and to be an effective use of time (3). Some of the free text comments from students echoed these points too. One student observed that ACJ gave them *"some ideas how to be creative in writing and for what to looking for (sic) when you have to judge the text".* In this excerpt this student identifies how ACJ of creative writing provided them with models of creative writing that helped to demonstrate to them what 'good ' looked like. This comment would certainly merit

further exploration if it was shared in a classroom setting, as there would be value in understanding further what creative ideas this student acquired from reading their peers work.

Moreover, and staying with the student comment above, there is a risk that we can identify in using ACJ as a form of peer assessment in that this student is taking the examples they are reading as models of good practice. While the scripts might be representative of good writing, there is a risk in any unmoderated sample of student work that students judging quality when they have not yet developed a discerning eye for good quality might reinforce poor quality or incorrect practices in their own writing. It is here that we can observe the potential value of using NoMoreMarking's judging results page (as seen in Method 1 and 2 for teachers) to provide insight into student judging performance, including an infit score for each student to show their agreement with one another. This process could be even more robust if judging was undertaken by students and teachers across the same sample, allowing for insight into agreement across all judges.

A theme emerging from the free text comments that students shared was the role of peer assessment as an enabler for self-assessment to take place. Students noted that through ACJ "*it was good to see other answers and then compare to your own one to see if you can make any improvement,"* and *"I also learnt what level I am compared to many of the students are (sic). I am at a good level."* This chimes with Marshall and Wiliam's (2006) assertion that 'peer assessment is one of the main vehicles to promote self-assessment', owing to the fact that 'seeing how someone else has tackled the same assignment helps pupils reflect on their own performance'

(2006:19). This is to be achieved through discussion with others about their writing, which enables them to 'gain insight into what is involved in good writing and thus...extends the range and scope of their repertoire' (Marshall and Wiliam, 2006:5). Method 5 explores the nature of student dialogue in discussing the judgement decisions they arrived at, the findings of which are presented below. This is another theme that is explored in greater depth in Chapter Six.

## Findings from Method 5 - semi-structured interviews with students

The findings from this method are reported thematically. The sections below each represent an emerging theme that was evident across at least one interview with students in discussing their use of adaptive comparative judgement in considering the quality of different creative writing scripts. These themes are: the concept of flow as an indicator of textual quality, engaging with the text aesthetically, and adaptive comparative judgement as a tool for self-reflection.

### The concept of flow as an indicator of textual quality

One of the most surprising findings that emerged from the interviews with students was the consistent references made to textual flow. Students indicated that they felt flow was important as an indicator of textual quality, a reference that was also made during the interviews with teachers in method 3.

There were statements that spoke of textual flow as important for the enjoyment of a text:

***Interviewer:*** *Is there anything in there that's more important than anything else?*

***Student D:*** *The timeline of the events flows (sic), how the events flow...because if you get too complicated then you get lost and you don't know what's happening, and so that makes the text less enjoyable.*

Other contributions from students spoke of how a text flowed was a deciding factor in how successful a piece of creative writing, as below:

***Student L:*** *Well, firstly, I went through all of them in order and read through the ones that caught my eye, or that made the most sense to me, in a punctual way, a descriptive way, in the flow of the story.*

---

***Student M:*** *This text (referring to one of the texts from the sample) had lots of description and flowed well*

***Interviewer:*** *what do you mean by flowed?*

***Student M:*** *like, it carried on, it didn't have a huge chunk missing out of it*

One student cited a particular example from an opening sentence from one of the

texts, contesting that the phrasing was awkward and ineffective as a result. They initially struggled to articulate this but again the concept of flow was employed to account for this, as detailed below:

***Student L:*** *when it comes down to most of them, they could have been worded better*

***Interviewer:*** *So, it's like a phrasing thing…*

***Student L:*** *yeah the ideas aren't bad, it's just the phrasing is off...the wording is off and something isn't clicking*

***Interviewer:*** *So what it is that's off, can you narrow in on it?*

***Student L:*** *It doesn't flow. For example, in the first line (reading from one of the texts) "the first time I saw a dead body was the one I killed. This was many years ago. This has happened and now I'm in prison for it. Here's my story". It's a bit...it could have been...it doesn't flow... (pausing)...I'm trying to think of a way I could have worded it.*

One of the group interviews with students saw the concept of flow discussed alongside some concerns over its seemingly intangible properties:

***Student A:*** *(referring to one of the texts) "there were flowing sentences...the whole thing flowed..."*

***Student J:*** *"it's flowy…"*

***Student F:*** *"But we won't get that in the article (sic)* because how it flows is going to change all the time"*

**clarified after the interview - referring to the exam*

In the above exchanges student F raises a valid concern over the concept of textual flow as being achieved through a multitude of deliberate and varied composition choices, and how this changes for each individual text. The concept of flow has already been discussed in some detail above in the findings from method 3, and the semi-structured interviews with teachers. It is a theme that will be explored in further detail in Chapter Five. In these explorations consideration will be given to addressing the justifiable concerns raised by Student F that flow, while clearly an important factor to consider in textual composition, is difficult to characterise. Efforts will be made to try and understand this concept of flow at a more fundamental level, so as to potentially offer some insight into how students might be supported in achieving it in their own writing.

## Engaging with the text aesthetically

Another theme that was evident from the interviews with students was evidence to suggest that students were judging the quality of creative writing through the use of an aesthetic lens. This can be seen as distinctly different to a technically-aligned

critical lens that might focus on spelling, punctuation, grammar and some of the more granular components of creative writing.

In the exchange below, two students explained how they enjoyed a twist one of the texts featured:

***Student K :*** *Which one had the better story line.*

***Student M:*** *Yeah, and which one had the better descriptions. And good use of language*

***Interviewer:*** *So let's explore that a bit further...so first of all when you say storylines, what does that look like to you?*

***Student K:*** *so, the ones that made me know what happened next, the ones that had a different kind of story*

***Interviewer:*** *could you give an example?*

***Student K:*** *so (referring to one of the texts from the sample) this one was good because no-one writes about a plane crash...it was interesting*

***Student M:*** *and unique, and made you want to read on*

***Interviewer:*** *what about it made you want to read on?*

***Student K:*** *it starts off like a normal plane ride, just a person on a plane, and then it's like oh, plane crash. It just happens*

***Interviewer:*** *did you see it coming - the crash?*

***Student K:*** *take away the plan and no (laughing). I just thought it was going to be a description, a journey.*

Another exchange featured students reflecting on the importance of texts grabbing your attention:

***Student G:*** *There was one I read, I just read the first paragraph and I was like "wow", you know.*

***Interviewer:*** *What was the "wow" moment?*

***Student G:*** *It was just the descriptive...the language devices employed, and stuff like that*

***Student B:*** *and I think the way it starts too, the attention*

***Student G:*** *yeah, your attention*

This aesthetic engagement is important for several reasons. Firstly, there are indications that students were adopting the perspective of a reader, and by doing so were formulating and cementing an understanding of textual audience. This is a crucial component not only of creative writing, but of other aspects of the GCSE English curriculum too. Further to this, aesthetic engagement reveals that students were considering textual quality in respect of the whole of the text, rather than its constituent parts. Students K and M spoke of how the twist in one of the texts made then want to read on. While they did not explicitly state it, what they alluded to is how the structure of a piece of creative writing can interest a reader. But in order to successfully understand this as a device that can be employed in narrative

composition, it could be argued that there is a need to appreciate the device in context from an aesthetic, rather than technical, perspective. This can only be achieved when the text is taken as a piece of creative writing to be read, engaged with and enjoyed, rather than as a technical demonstration of writing ability. In situations in which students peer assess adopting methods that focus more on the latter Marshall and Wiliam (2006) warn that 'they can oversimplify the complexity of a good piece of writing...and so misrepresent for the pupil what needs to be done.' (2006:15).

## Adaptive Comparative Judgement as an enabler of self-reflection

The final emerging theme in this section to discuss is evidence that suggests that ACJ can serve as an effective device to facilitate self-reflection in students focusing on what makes good writing.

One student commented in their interview that the ACJ process had led to self-dialogue:

***Student G:*** *when I was reading it you picture yourself there. There was one talking about the clouds making animals and I'm just there trying to visualise it to see what they're writing about*

***Interviewer:*** *and was that one a good text?*

***Student G:*** *I asked "why didn't I come up with that myself?"*

In other exchanges there was evidence of students reflecting on the process, and at the same time making meaning through conversation:

***Student K:*** *this one had a load of description, but I thought it was too much*

***Interviewer:*** *so there's such a thing as too much descriptive writing?*

***Student K:*** *...just pages of descriptive writing*

***Interviewer:*** *so what's it lacking then, if all it's doing is describing, what is it missing?*

***Student M:*** *it just stays in one place*

***Interviewer:*** *so a strong story is one that moves?*

***Student K & M*** *(in unison): yeah*

The choice in designing the interviews to be semi-structured was vindicated through interactions like the above, which allowed me to explore arising thoughts and themes in greater depth and clarify points that students had made. The focus of the judgement in an ACJ approach to assessment to always be one of comparison was successful in leading to productive and meaningful dialogue between students. This dialogue was enriched through the presence of multiple points of reference in any given judgement scenario. To exemplify, the conclusion arrived at my students K & M on the importance of a story that *"moves"* rather than remains in one location was realised by virtue of having others texts they could refer to in acting as a counterpoint, that modelled to them what good looked like in this context. Whether they would have arrived at the same conclusion if they had not been scrutinising

texts in a comparative fashion is not possible to know, but one ventures that they would likely not have.

## Concluding remarks

The findings reported in this Chapter provide valuable insights into the research questions that underpin this enquiry. A summary of these findings is featured below:

Methods 1 & 2 determine that teachers were able to successfully use adaptive comparative judgement in assessing creative writing script quality, and that their judgement decisions were largely in agreement with one another despite not calling upon mark schemes or assessment standards to inform these. Methods 1 & 2 also suggest that there is a correlation between time taken per judgement and how consistent judgements were.

Findings from Methods 1, 2, 3 in tandem indicate the existence of a relationship between a judge's agreement score and their experience in teaching GCSE English, pointing to professional learning as being crucial in becoming an effective indicator of the ability to judge textual quality. Along the same thread, Method 3 determined that formal professional learning in the teaching and assessing of GCSE English in Further Education settings for the teachers interviewed often lacked specificity and focus on pedagogical matters. Conversely, teachers felt that dialogue with peers was vital in sharing practice and setting standards. In addition, Method 3 highlight how teachers drew upon their tacit understanding and instinct when judging text quality,

referencing concepts such as creativity, textual flow, and to what extent a text engaged them.

Method 4 identifies that the majority of students found adaptive comparative judgement to be an activity that they learned through participating in, a way of furthering their understanding of the subject and a valuable use of time. Methods 4 & 5 in tandem uncover how students undertaking peer assessment through ACJ resulted in instances of self-assessment occurring. Method 5 indicate that students, much like teachers in Method 3, successfully drew upon their tacit understanding of what ‘good’ looked like when determining textual quality, and articulated quality indicators through metaphor.

# Chapter 5 – Discussion

## Introduction to the chapter

This Chapter provides a more extensive discussion of the findings, and other themes that have emerged from this research. It positions these alongside theoretical models and research centred on assessment practice, professional learning and comparative judgement. The intention here is twofold. Firstly, to provide a broader context through which the findings of this study can be understood. Secondly, to consider the contribution to knowledge emerging from this enquiry.

Frequent references are made throughout this Chapter to important theoretical works and research that were cited in Chapter Two as the groundwork for this enquiry was being laid. In addition, references are made to work that was previously unreported is now deemed relevant in view of what this enquiry has uncovered. References of this nature can be considered to be representative of new, emerging or unforeseen trends that were not initially anticipated, but are valuable nonetheless in seeking to address the research questions posed in the enquiry.

This Chapter is organised into two main sections. It begins by examining in greater depth the findings from the adaptive comparative judgement trials conducted with teachers. This is in part realised through comparisons of results from other ACJ trials conducted across a range of settings that lead to a consideration of results from this research and what can tell us about using ACJ as an approach to assessment in Further Education settings. The second section focuses on tacit knowledge. The

centres more closely upon tacit knowledge a key component of assessment practice in English. Firstly, findings from this study are compared with those from others within the field. Following this, findings reported in the previous chapter are re-considered and analysed through the theoretical lens of conceptual metaphor theory (Lakoff & Johnson, 1980).

### Researcher positioning in the discussion of findings

The discussions below represent a more in-depth examination of some of the significant findings reported on in Chapter Four. It identifies and explores underpinning themes that are mapped to theoretical understandings and other empirical research so as to present a coherent and authentic account of what new knowledge this research has uncovered. These links to theory and other research are made intentionally, so as to anchor this research in a situated context alongside similar and different work in the domains of educational research and theories of assessment practice. As a practitioner researcher there is a risk during this phase of the research that my beliefs, values and experiences might influence the nature of the analysis that follows. It is hoped that by sketching a broader picture of what the emerging findings from this research mean in a wider context with reference to contemporary works that such risks are mitigated.

## Adaptive comparative judgement

This first section of this Chapter seeks to examine in greater depth and detail the findings gathered through the adaptive comparative judgement trials with teachers.

The aim here is to locate a context within which these findings can be better understood, and to seek to answer questions that arose when these findings were initially reported in Chapter Five.

The trials yielded findings of several varieties, comprising teacher agreement with one another (infit), how many judgements were made during the allocated time frame (local) and the median time per judgement (median).

## Reliability

Both trials also received an overall reliability score. It is on the matter of reliability that we begin these discussions.

**Method 1 ACJ reliability:** 0.95 Scale Separation Reliability
**Method 2 ACJ reliability:** 0.93 Scale Separation Reliability

These results from methods 1 and 2 evidence a very high degree of reliability. Research in which a Scale Separation Reliability (SSR) measure has been applied in the use of rubric based approaches to marking demonstrates a significant lower figure for reliability. Doğan & Uluman (2016) identify a SSR score of 0.60 across a sample of 82 students' written work that had been double marked through application of a marking rubric. While an isolated study, it is evident that disparities across markers in this instance were vast when use of a marking rubric was applied. The lack of reliability in the use of rubric based mark schemes is also evident in the findings from *NoMoreMarking* (2017) a survey of research into GCSE English

reading responses reported in Chapter Two, and the critical incident that underpinned the discussion of the problem on which this enquiry is based in Chapter One. While no SSR score is available for these two examples, they represent clear deficiencies in assessment reliability that are evident even without a numerical value attached.

If we are to take that adaptive comparative judgement offers a more reliable approach to assessment than conventional rubric-based processes, then we can consider what the reliability findings from this enquiry translate to in the wider context. Do these scores of 0.95 and 0.93 correlate with the findings from other research in which ACJ has been used? And subsequently, what do these comparisons tell us about ACJ in the manner it was used in this enquiry? Bramley (2015) reports on the findings from thirteen published research studies in which adaptive comparative judgement was employed as a way of determining the quality of student work. These studies span a range of subjects and were largely conducted in UK Secondary Schools. These findings are presented in the table below.

| Study | Adaptive? | What was judged | #scripts | #judges | #comps | %max | #rounds | Av. # comps per script | SSR |
|---|---|---|---|---|---|---|---|---|---|
| Kimbell et al (2009) | Yes | Design & Tech. portfolios | 352 | 28 | 3067 | 4.96% | | 14 or 20 bimodal | 0.95 |
| Heldsinger & Humphry (2010) | No | Y1-Y7 narrative texts | 30 | 20 | ~2000? | | | ~69 | 0.98 |
| Pollitt (2012) | Yes | 2 English essays (9-11 year olds) | 1000 | 54 | 8161 | 1.6% | 16 | ~16 | 0.96 |
| Pollitt (2012) | Yes | English critical writing | 110 | 4 | (495) | (8.3%) | 9 | ~9 | 0.93 |
| Whitehouse & Pollitt (2012) | Yes | 15-mark Geography essay | 564 | 23 | 3519 | 2.2% | (12-13) | ~12.5 | 0.97 |
| Jones & Alcock (2014) | Yes | Maths question, by peers | 168 | 100,93 | 1217 | 8.7% | N/A? | ~14.5 | 0.73<br>0.86 |
| Jones & Alcock (2014) | Yes | Maths question, by experts | 168 | 11,11 | 1217 | 8.7% | N/A? | ~14.5 | 0.93<br>0.89 |
| Jones & Alcock (2014) | Yes | Maths question, by novices | 168 | 9 | 1217 | 8.7% | N/A? | ~14.5 | 0.97 |
| Newhouse (2014) | Yes | Visual Arts portfolio | 75 | 14 | ? | ? | ? | 13 | 0.95 |
| Newhouse (2014) | Yes | Design portfolio | 82 | 9 | ? | ? | ? | 13 | 0.95 |
| Jones, Swan & Pollitt (2015) | No | Maths GCSE scripts | 18 | 12,11 | 151,150 | 100% | N/A | ~16.7 | 0.80<br>0.93 |
| Jones, Swan & Pollitt (2015) | No | Maths task | 18 | 12,11 | 173,177 | 114% | N/A | ~19.5 | 0.85<br>0.93 |
| McMahon & Jones (2014) | No | Chemistry task | 154 | 5 | 1550 | 13.2% | | ~20 | 0.87 |

*Table 1: Design features and SSR reliability results from published CJ/ACJ studies, taken from Bramley (2015)*

Evident in the table is that the SSR values have been high or very high in published work where CJ or ACJ has been used as an alternative to marking. In these the majority were in excess of 0.9 and only was below 0.8. In view of these, the SSR results from this enquiry represent very high degrees of reliability, even in the context of ACJ which is in itself has been demonstrated to be a highly reliable approach to assessment.

There are two points of interest in respect to the findings from studies that Bramley reports. Firstly, many of the research studies were larger in scale when compared with this enquiry, featuring more scripts and judges. This can perhaps account for the relatively high reliability scores achieved in this enquiry when compared with these studies, in that a smaller pool of scripts and judges resulted in greater homogeneity in their judgements. Secondly, and perhaps more significantly, many of the studies reported in the table represent research that has been conducted on using ACJ approaches in subject areas other than English. Only two of the thirteen are explicitly identified as focusing on this subject. Comparisons between results from this enquiry and these two studies in respect of reliability reveal similar results: Pollitt (2012a) identifies reliability values of 0.96 and 0.93 respectively. But the use of ACJ approaches to explore assessment practices within other subjects exemplifies the challenges that teachers spanning multiple disciplines face when seeking to interpret assessment standards. The existence of CJ/ACJ research focused on assessment practices in subjects including maths, science, geography and design, points to the need for research to seek to address perceived issues in assessment

such as those included in this enquiry, namely the inherent subjectivity in assessment standards. For example, open questions in which students need to demonstrate an opinion or argue a position are always going to be difficult to align to standards, regardless of the subject. What we can recognise here is the ubiquity with which English, through the application of language, underpins all other subjects.

So, what can we take from the reliability values across these studies? The consistently high reliability ratings found in this study appear to indicate that there exists within subject disciplines, agreeable forms of understanding regarding what makes good quality work that teachers can successfully recognise and draw upon when forming a judgement, regardless of the topic or subject matter. The challenge of interpreting standards appears to be a prevalent one, certainly in respect of demonstrations of knowledge and understanding that are relied upon in some way by language. Circling back to this the findings of this enquiry and the reliability values determined in methods 1 and 2, with such high agreement scores across both methods, it appears as though a mutually shared tacit understanding is being drawn upon by teachers when they were forming their judgements. As discussion of the findings from this study progresses, the importance of interrogating this concept of a shared idea of what good work is comes into view.

We have ascertained above that there is a shared idea between the teachers that featured in both methods 1 and 2 of this enquiry of what makes good creative writing. Before we move beyond discussions around the reliability of assessment decisions ascertained in this enquiry, there is a need to consider the very nature of a reliability value that is derived from agreement between judges. Reliability by

definition suggests precision and objectivity, but what we can note is that the reliability values derived from this enquiry represent an agreement between teachers on compared creative writing scripts. With no external anchor, the reliability rating solely represents teachers' agreement with themselves. This is not necessarily problematic on its own. Agreement is certainly preferential to disagreement, and points to the notion of a shared construct of what comprises good work. But there are questions to be asked about the validity of the judgements being reached. What if all teachers hold an equally misrepresented understanding of what makes good creative writing, and the reliability value merely represents an alignment of misunderstandings?

This question is all the more pertinent when considered alongside the matter of domains of knowledge and understanding that exist in localised environments. To exemplify, Morrison et al. (1994) identify systematic differences between grammar school and secondary school teachers in what they valued in students' work. We can assume that these differences in what was deemed valuable came about through the assimilation of teachers into the values, norms, standards and expectations relative to each school, and that this had an impact upon what they deemed to judge to be valuable accordingly. As Wiliam (2016) notes, 'the rank order emerging from comparative judgement scoring depends on a relatively coherent community of interpreters.' The word 'coherence' can be interpreted to represent agreement across a small group of teachers, as seen in the teacher groups involved in method 1 (seven teachers) and method 2 (five teachers). But these groups do not even represent the entire teaching team for GCSE English within their institution. In the

context of the study conducted by Morrison et al. (1994) it is clear that disparities in coherence can exist on a national scale.

Rather, it could be argued that ‘coherence’, in the context of Wiliam’s quote, should be taken to represent agreement across a community of interpreters and that that agreement should be as comprehensive as possible. In turn this means that in order to be representative, this community should be diverse, in respect of experience, working environment and values set. Through this, the coherence that the community strives towards will be representative of, and enriched by, a broad, socially constructed understanding and knowledge of what comprises a good standard of work in a subject or topic.

So, what does this mean for comparative judgement, and this enquiry? What we can note is that there was clear agreement between the teachers in both ACJ trials about what made good quality work. However, using the term reliable to describe these judgements necessitates some caution, in that it could imply that they are reliable in respect of an externally located, nationally defined/prescribed standard of what we mean by good quality work. It is possible that the two conceptions of good, that the teachers in methods 1 and 2 identified, and what awarding bodies and standards verifiers for GCSE English define, overlap. Indeed, they might well be the same. But it was not the intention of this enquiry to standardise assessment judgements formed through ACJ against mark schemes and assessment standards. This remains an area of potential future enquiry. As a result, use of the term reliable in describing the assessment judgements in this enquiry comes with a caveat: these results were reliable, in that they demonstrate a high degree of agreement between judges in

respect of their personal interpretations of what good quality creative writing is in the context of GCSE English.

## The role of experience

The matter of personal interpretations leads us to consider the role that individual judges had in forming this set of judgements. Personal interpretations of what good quality creative writing looks like is ultimately tempered through an individual's experiences. It is impossible to detach the role of experience when considering the consistency of judgement that each teacher made. It is here that we can seek to answer one of the questions posed in Chapter Five: Is there a correlation between how experienced a teacher is in teaching GCSE English and the consistency of their judgements compared with their peers?

The following discussion references the findings obtained during methods 2 and 3 working with teachers 8, 9, 10, 11 and 12. The working experiences of teachers 1-7 that participated in method 1 were not captured, and hence are not represented here. As reported in Chapter Five, from comparing teachers infit scores with their reported experiences and training profiles we can note that there is a positive correlation between how experienced a teacher is in teaching GCSE English and how consistent their judgements are compared with their peers.

These findings align with findings identified by other researchers focusing on the relationship between experience and judging consistency. Whitehouse and Pollitt

(2012) used comparative judgement to look at responses to an AS level geography exam, with markers that were more experienced in teaching the specification (A-Level geography teachers) and less experienced (GCSE geography teachers). What they found was that there was less consistency across the GCSE teachers than amongst A-Level teachers. Subject knowledge did not take precedence here, whereas familiarity and teaching experience with the specification being examined did. This suggests that tacit knowledge grows with familiarity with the course and specification, rather than with subject knowledge alone.

Teacher 12 provides us with an interesting case study, in that they were the least experienced of the five teachers, and at the time of participating in the ACJ workshop were working towards completing a teaching training qualification. Their experience of teaching GCSE English had to that point only comprised approximately three hours a week for the past eighteen months. This is in contrast to the other four teachers who had an average of over three years of full-time teaching of GCSE English. It is perhaps expected that teacher 12's assessment judgements might differ from other teachers, in that they have not yet established a comprehensive base of knowledge and experience on which to build these on. What we can take from this is that judgement practice is refined and developed a period of time and not simply acquired as a form of 'propositional knowledge'. Whereas propositional knowledge constitutes 'know-that', tacit knowledge represents the 'know-how' (Winch, 2010), in which knowledge is applied to a practice as with the judgement practice that is applied during comparative judgement.

## Judgement duration

Chapter Four reports on the median times that judges across methods 1 and 2 recorded when comparatively judging. It was evident from the findings that there were vast disparities in how long each judging decision took, depending on the judge. The questions raised in that chapter centred on the relationship between judgement reliability when compared with duration per judgement, and if we could consider there to be an optimal time that judges should spend on average when judging scripts.

Before exploring these questions, it is worth noting the importance of considering the role of duration per judgement. We can recognise the value of having an opportunity to explore the duration that teachers took when assessing work. In conventional marking approaches no such measure is available unless specific conditions are arranged in advance, and it is hard to see how such conditions would not interfere with the teacher in such an arrangement. Because ACJ through the *NoMoreMarking* software calculates and reports on duration per judgement as part of its design we are in a position to consider the significance of how long teachers took on average to judge the better script when presented with two options. This is valuable insight into assessment practice, and represents a new opportunity that was otherwise not accessible to teachers before the development of technologies that facilitate its possibility in recent years. This recent development can perhaps account for the relative paucity of literature and research reporting specifically on the role and significance of duration per judgement in comparative judgement assessment scenarios.

The first question posed above centres on the relationship between judgement reliability compared with duration per judgement. In respect of the findings gathered from methods 1 and 2 it is difficult to note a correlation between reliability and duration taken, beyond the observation that taking longer per judgement leads to a greater reliability. Perhaps the most significant finding from this measure is the significant disparity in duration taken per teacher. Teacher 3 in method 1, for example, had a median time of 3.3 seconds per judgement. They had the lowest judgement duration across all teachers that reported an infit value of 1.0 or under. In contrast, Teacher 4 in method had a median time of 149.6 seconds per judgement. Chapter Four listed some possible causes as to why this disparity might be so significant, including the possibility that this teacher chose to re-read every script during each combination. In respect of the findings reported through methods 1 and 2 it is not possible to draw conclusions on the significance of an individual's duration per judgement and how reliable their judgements were with any confidence.

This is not to say they are not valuable findings to be considered in relation to the group as a whole. We can note the difference in research design between methods 1 and 2. Method 1 was launched with a short training and workshop session and followed with an instruction for teachers to complete their outstanding judgements independently. Method 2 comprised one workshop session and a set duration in which teachers completed judgements within this. In both methods no specific instructions were given as to how long it should take to arrive at a judgement. Teachers were given freedom in this regard to take as long as they felt they required. Method 1 demonstrated a far greater disparity in median judging time compared with method 2, which by comparison saw a far greater correlation.

The differences in research methods here are significant. We can tentatively assert that the workshop and subsequent individual allocation of scripts led to a greater disparity in the undertaking of judgement practice for teachers involved in method 1. In contrast method 2 yielded far more consistent judgement durations. In Chapter Four it was suggested that this might be due to the shared environment that all teachers inhabited when completing the ACJ trial. What this points to is how conformity to community established norms of practice, such as how long it should take to judge the better of two scripts, is influenced by the environment in which such practices are subconsciously agreed, maintained and perpetuated. A point to consider here is how an individual might disrupt any pre-established norms for better or worse. For example, if a judge with an extremely low median time from method 1 had joined the method 2 workshop, would the average duration be affected? Or would they conform to the group's pre-established norms? And ultimately, would this have any bearing on the reliability on the judgements made? This enquiry is not in a position to report on the impact of such conditions, but this remains a line of enquiry that would be worth pursuing.

It is here that we can locate considerations of the second question carried over from Chapter Four, regarding the suggesting of an optimal time that judges should spend on average per script. The findings in methods 1 and 2 do little to suggest there to be an optimal time that should be spent per script. What they do indicate, however, is how undertaking adaptive comparative judgement in a group setting while still working individually forming your own judgements can be beneficial in respect of normalising specific elements of judgement practice. On the matter of an optimal

time for judging, questions remain as to if it is even desirable or viable and in the interest of effective assessment practice. Chapter Four observes the pragmatic advantages that determining an optimal time per judgement would offer. But an optimal duration would only ever be a heuristic at best, suffer from trying to apply to all judges rather than any specific one, and only apply to the task students had completed in that sample. The discussions that follow this section focus on assessment practice, and chart some of the challenges we face if it is positioned as a procedural activity.

## Tacit knowledge

The next section of the chapter is centred on examining in greater detail what it is that is helping teachers form their assessment judgements. To frame this discussion, we can firstly look again to Sadler's (1989) conception of 'guild knowledge', on which he writes:

> 'Teachers' conceptions of quality are typically held, largely in unarticulated form, inside their heads as tacit knowledge. By definition, experienced teachers carry with them a history of previous qualitative judgments, and where teachers exchange student work among themselves or collaborate in making assessments, the ability to make sound qualitative judgments constitutes a form of guild knowledge' (1989:126).

This offers a sound starting point for this discussion as it is presented within the domain of educational assessment. In guild knowledge, Sadler is attempting to

articulate how teachers develop a tacit understanding of good quality over time, as a result of repeated exposure to different examples of work through collaboration with other professionals. This is not a distinct concept that is located aside to some of what has already been discussed above. As Marshall (2011) observes: “what James Britton called impression marking was similar to Sadler’s guild knowledge [...] what Sadler calls ‘their essential fuzziness’ and perhaps James Britton calls an impression, for others is the term ‘judgement’’ (2011:26-27).

Guild knowledge by its very name draws on the idea that this knowledge exists in a practice-oriented community. Guilds of the middle ages, centred on craft, were located in workshops that served as a ‘productive space in which people deal face-to-face with issues of authority [...] In a workshop the skills of the master can earn him or her the right to command, and learning from and absorbing those skills can dignify the apprentice or journeyman’s obedience’ (Sennett, 2008:54). For Sennett there is an incumbent need for a practice to be guild-oriented, so to address that which cannot be achieved through individual autonomy. Issues of authority are a part of this. Medieval guilds addressed this through the master craftsman, ‘a superior who sets standards and who trains’ (ibid:54). The role of a master in both setting standards and training others in recognising and realising these standards in their own practice is important. He notes that well-crafted institutions will favour the sociable expert; the isolated expert sends a warning signal that the organisation is in trouble (ibid:246).

Tracing back to Sadler’s guild knowledge, we can recognise that the purported standards of any guild are owned by its members and set, sustained and developed

in the practice the guild represents. The role of the master craftsman, as in medieval guilds, is perhaps not a viable construct in the context of a team of teachers, but we can recognise some parallels between a master craftsman and what we might regard as more senior members within a teaching team. These include their relative experience, the quality of their work as regarded by their peers, and the responsibility in training less experienced and new members. The relationship between a master and apprentice is perhaps more recognisable when considered in respect of a teacher and the students they teach. We can note here the value in examining the content of what the guild knowledge of teachers and students comprises of in respect of this enquiry, in an attempt to determine what this form this tacit understanding of good quality takes. This will include excerpts from those ostensibly more adept in judgement practice, that is to say the teachers, and those that are currently apprenticing in the craft, students.

## Tacit knowledge and judgement practice: comparisons with other approaches to the assessment of writing

Chapter Two reports on different approaches to the assessment of writing that have been explored over the previous seventy years or so. These included the works of James Britton (1950) into the marking of imaginative compositions, Wiliam's (1994, 1996, 1998) use of construct referencing, and D'Arcy's (1999) adoption of interpretive responses to student writing. The discussion which now follows revisits some of this work, and other work pertinent here, and considers parallel findings identified in this

enquiry, and what it tells us about judgement practice in the context of English and creative writing.

**Interpretive response judgement**

Britton's (1950) work that led to the publishing of the *Report on the Meaning and Marking of Imaginative Compositions* was undertaken with the aim to look at how English might be assessed more holistically and reliably. Seven individuals took part in this project, and were first asked to think of criteria they would use to mark compositions. From this they selected 'two items which, between them, seemed to cover the greater part of what we meant my imaginative composition. These were a) pictorial quality and b) creativeness' (ibid:2). Pictorial quality represents the way in which a writer creates an image in words and describes something in detail, while creativeness is defined thus: 'To what extent is what the writer has written new, original or individual?' (ibid:2). As noted in Chapter Two, marking through these criteria led to different interpretations. Although interpretations were diverging, we can recognise here the intention of Britton and his colleagues to give credence to the importance of creativity and originality in the judging of text quality, a quality that teachers involved in this enquiry also identified as important.

Britton's first attempt led to the development of a second, more refined approach to how judges were to interpret co-constructed criteria. In a second cycle of the judgment trial, teachers were asked to write a 100-word piece and changed the criteria again, asking for:

> ‘1) General impression (by your own personal method; by impression rather than by analysis in search of particular characteristics).
> 2) To what extent can the reader experience what is presented (i.e. see, feel, hear etc.)
> 3) Originality of ideas. To what extent is the writer’s view of the subject distinctive (i.e. as compared with the ideas of the group as a whole.).
> 4) Feeling for words. To what degree does the writer use words a) strikingly AND b) effectively?’
>
> (ibid:3)

A common theme evident running through each of these criteria is the interpretive nature of each. Teachers were explicitly asked to consider what they felt through impression. Criterion 1) explicitly states that no analysis in search of particular characteristics should take place. We can recognise how these criteria were constructed in a way to encourage a more holistic judgement of quality. Significantly, the criteria place an important emphasis on teachers using their ‘gut feeling’, and what we can understand to be tacit knowledge, in judging text quality. The results from this second attempt were much akin to the first, in that judges still disagreed with one another as a result of varying interpretations of the standards.

There are two things that we can tentatively conclude from this. The first is that interpreting criteria, even criteria that encourages interpretive responses that enable teachers to draw on their tacit understandings of what makes good quality work, still encounters difficulties in assessment validity when adopted in an absolute referencing assessment scenario. The second pertains to a consideration of community-centred ownership of standards. In this trial Britton was focused on

determining the reliability of the judgements that teachers reached. The research design was centred on ensuring this reliability, and the teachers involved completed both individual marking of each criterion and repeat marking in attempts to maintain this. We can appreciate that with this Britton sought to prove this assessment method to be viable for wider practical application. But what was possibly lost as a result of the trial being designed in this manner was that teachers involved did not have the opportunity to share their interpretations and attempt to align their own internal standards of quality. In the context of the trial, the validity of judgements was neglected in favour of their reliability.

What is striking in respect of the work of Britton and his colleagues is how the explicit criteria they developed in this trial, with forethought and a deliberate focus on forming an interpretive response to text quality, align with some of tacit standards that teachers involved in this enquiry appear to have drawn upon when doing the same thing through comparative judgement. The difference is that no explicit criteria were given to teachers in this enquiry, beyond the instruction to 'select the more proficient text'. Teachers in this enquiry framed the text as a whole when considering quality, echoing '1) general impression criteria', spoke of asking themselves "did it grip me?", pointing to a tactile response linked to '2) reader experience' criteria, and the role of creativity we can align with '3) original response' criteria (ibid:3). Interpretive responses, therefore, are an important facet of judgement practice.

**Construct referencing**

Dylan Wiliam's (1994, 1996, 1998) work into construct referencing took a different approach to assessment that set it apart from criterion referencing. In summarising this, Marshall (2011) states 'In essence, when teachers of English award a grade to a text, they draw on a construct of what they think that grade looks like, based on their previous encounters with work of a similar standard - very like Britton's impression marking' (2011:27). Central to this a community of interpreters made up of teachers. Through constant debate with other teachers from different schools as to what construct best applies in any given instance, a professional discourse emerges. This in turn leads to shared meaning across all the community of interpreters. On this, Wiliam (1998) notes:

> 'The innovative feature of such assessment is that no attempt is made to prescribe learning outcomes. In that it is defined at all, it is defined simply as the consensus of teachers making the assessments. The assessment is not objective, in that there are no objective criteria for a student to satisfy [...] the assessment system relies on the existence of a construct (of what it means to be competent in a particular domain) being shared by a community of interpreters' (1998:6)

We can recognise the parallels between Wiliam's approach to construct reference marking and comparative judgement. Both forms of assessment dispel the use of criterion-based standards that are required to inform the judgement being reached. Rather, they both advocate the use of an interpretive and tacit understanding of what good looks like, as determined by the community. What the approach adopted by Wiliam through construct referencing does differently to comparative judgement is its

openly dialogic nature. It invites dialogue that serves to create a shared understanding across the community. It could be argued that adaptive comparative judgement, as facilitated through the use *NoMoreMarking* or similar software, also enables this. The adaptive nature of the assessment design in ACJ means judges negotiate through and eventually produce an ordered list that is reflective of the quality of texts within a sample. What ACJ does not facilitate is the dialogue with other members of the community *during* the act of judgement. This can only take place afterwards.

What ACJ does enable, however, is self-dialogue. This is because of the comparative nature of the assessment design. Dialogue with other members of the community through which standards are challenged, debated and shared through methods as evident in Wiliam's construct referencing is critical, and we can understand that a similar process is taking place when comparative judgements on quality are made. Dialogue, whether with one's self or with an external party, is about more than putting one's own view across. It is about meaning making. For Sennett, (2008) 'to do good work means to be curious about, to investigate, and to learn from ambiguity [...] craft negotiates a liminal zone between problem solving and problem finding'. Several teachers in this enquiry articulated how they were asking themselves questions when considering the quality of texts, and we can recognise this to be possible evidence of them undertaking a self-dialogue as they negotiated the problem posed by the comparative pairing of texts, in what Sennett terms the negotiation of liminal space.

**Meaning making through metaphor**

As observed in Chapter Four, one of the most prominent recurring findings across both the teacher and student interviews was reference being made to the concept of 'flow'. The frequency with which it was referenced makes it an important and interesting phenomenon to explore in some depth. So, what can we understand to be meant by the use of the word 'flow' in describing a creative writing text? What does it mean for *'events to flow',* for a story to *'flow well',* or for it to feature *'flowing sentences'?* Conversely, what can we understand of a text that *'doesn't flow'?* In order to answer these questions, we first need to explore in greater detail the construction of meaning that stems from the use of figurative language and metaphor in describing this quality.

We can start here with an expansion of the very concept of metaphor and its function in everyday language. Metaphor, as Johnson (1980) observes, "is no longer confined to the realm of aesthetics narrowly conceived - it is now coming to be recognised as central to any adequate account of language and has been seen by some to play a central role in epistemology" (1980:3). We can trace understanding of metaphor as a device for meaning making back to Ancient Greece, with Aristotle in his *Poetics* observing that:

> 'Metaphor consists of giving the thing a name that belongs to something else; the transference being either from genus to species, or from species to genus, or from species to species, or on grounds of analogy' (1457)

While the fundamental meaning of metaphor as proposed by Aristotle, as a means through which meaning is transferred from one vessel to another through comparison, has remained constant, what we can understand about the way that it is employed has evolved. Lakoff and Johnson (1980), in their seminal *Metaphors We Live By*, extend this far beyond the aesthetic, metaphor is 'pervasive in everyday life, not just in language, but in thought and action' (ibid:3). They contend that metaphor underpins our ordinary conceptual system, and in turn plays a central role in defining our realities. The term *conceptual metaphor theory* is coined to account for this.
So, what can conceptual metaphor theory tell us about the use of figurative language teachers and students used to describe what determined a specific text to be good quality? Lakoff and Johnson's belief is that use of metaphor in language results in the accessing of 'metaphorical entailments' that act as a reference point for the described vessel to be understood in relation to (ibid:106). As such, the entailments that are being accessed in a metaphor can provide valuable insight into an individual's conceptual understanding of this. Further to this, they contend that 'a given metaphor may be the only way to highlight and coherently organise exactly those aspects of our experience' (ibid:156). In other words, our understanding of the world is made possible through metaphor. With our intention to better understand the tacit knowledge of teachers and students, and the significance of them articulating what good quality work is through metaphor, we can recognise the value in examining the metaphorical entailments at the root of the word 'flow'.

In order to do this, we need to expand on the systematicity of the metaphorical concept entailed within the word 'flow'. Lakoff and Johnson demonstrate how we can

go about doing this by citing the conceptual metaphor of 'ARGUMENT IS WAR'. They reference the below examples as drawing on this conceptual metaphor:

> 'Your claims are *indefensible'*
> He *attacked every weak point* in my argument'
> His criticisms were *right on target*
> I *demolished* his argument
> You disagree? OK, *shoot*!'
>
> (ibid:4)

The systematicity of this metaphor draws on the notion that there are certain things that we tend to do and not do in arguments. It is noted that 'it is no accident that these expressions mean what they mean when we use to talk about arguments' (ibid:7). They argue that without such a metaphor we would not know what an argument *is.* To make the point, they ask us to imagine a world in which an argument is a dance. This calls on a visual depiction of argument as something far out of the ordinary.

So how might the systematicity relate to the metaphor of 'flow'? We can recognise that there are three slightly different applications of metaphor in the teacher and student their interviews. The first related to the whole text:

> *"This text...flowed well"*
>
> *"It doesn't flow"*
>
> *"the whole thing flowed"*

The second related to events:

*"The timeline of the events flows"*

The third was focused on a more sentence specific level:

*"there were flowing sentences"*

Despite the different applications of the *'flow'* metaphor, we can recognise commonalities in all three examples. In semantics the word flow describes a steady and continuous movement of some kind. As such, it is possible to assert that to say that a text, events or a sentence *'flowed'* points to the application of the same semantic meaning through a metaphorical construction. This is what we can understand to be what Lakoff and Johnson term a 'conduit metaphor' (ibid:10), which tells us that words are containers for meaning and that writers and speakers are containers for words. In essence, according to the flow metaphor writing emerges from within you.

Inherent in the conduit metaphor is the idea of movement, representing the transferral of meaning. When reading a text our eyes move over the words, and through this act we form meaning to comprehend what is written. If when reading a text, we encounter a word, phrase, or other feature that does not seem to fit in some way, or that otherwise breaks any immersion we had in the text, this steady and continuous transmission is impacted; the 'flow' of the text might slow, or come to a stop altogether. The more proficient the text the more efficient the transmission of meaning from words to us as readers. If a text is of poor quality this transmission of

meaning might slow down, or cease altogether. What we can observe in this example is the use of metaphor to describe something nebulous (how we form meaning from words on a page) through comparison with a more common relatable example (the movement of objects into a container).

We can note a significance from the data gathered in interviews at the frequency at which the 'flow' metaphor was employed by both teachers and students in describing what was representative of good quality creative writing. This points to a shared understanding across multiple individuals of this metaphor, its function and wider meaning. On this, Lakoff and Johnson note that 'metaphors may create realities for us, especially social realities. A metaphor may thus be a guide for future action. Such actions will, of course, fit the metaphor. This will, in turn, reinforce the power of the metaphor to make experience coherent. In this sense metaphors can be self-fulfilling prophecies' (1980:156). In other words, metaphors of this kind can enjoy a kind of ubiquity of use that means they become self-perpetuating. As the pervasiveness of the metaphor increases, the likelihood of it being employed increases too within a community. This chimes with the aforementioned concept of guild knowledge that comprises a tacit, community-owned understanding of what good quality looks like. The findings from this enquiry's interviews suggest that 'flow' is an important textual quality indicator in creative writing.

The question of how conceptual metaphor has been used to convey meaning through the example of flow has been addressed, but the question of why still remains not worthy of further explication. We can note that metaphor is often employed to ascribe meaning to ideas, objects and activities that we otherwise might

not be able to otherwise fully understand without a comparative frame of reference. But we can locate a more fundamental challenge faced by language as it seeks to make sense of our reality that is pertinent here. Lakoff and Johnson contend that the heart of the objectivist tradition in philosophy comes directly out of the myth of objectivism: the world is made up of distinct objects, with inherent properties and fixed relations among them at any instant (ibid:210). They advance that metaphor provides evidence against this perspective. In summarising their contestation, they note that:

> 'The objectivist philosophy fails to account for the way we understand our experience, our thoughts, and our language. An adequate account, we argue, requires:
> - Viewing objects only as entities relative to our interactions with the world and our projections on it
> - Viewing properties as interactional rather than inherent
> - Viewing categories as experiential gestalts defined via prototype instead of viewing them as rigidly fixed and defined via set theory' (1980:210)

In essence, meaning is relative rather than fixed. The idea of properties being interactional rather than inherent is in line with some of the challenges that have been previously addressed in this Chapter as to the static nature of absolutist assessment standards. To create a universal standard that is to be adhered to can lead to misrepresentation, misinterpretation and fuzziness in meaning. But this is not to say that standards do not exist. Rather, they can be more effectively understood (that is to say, acted upon and articulated) when experiential gestalts, such as metaphor, are permitted to account for what good quality looks like.

We can look again to Zimmerman's (2016) concept of 'fusion of horizons' as a hermeneutic device through which individually constructed meanings are shared and communicated as relevant here. As we have already identified, in comparative judgement assessment scenarios in which teachers and students are presented with the choice to judge the better of the texts, judgements are commonly made at a whole-text level, drawing on a holistic impression of quality. As such, attempts to articulate this require indicators that reflect the holistic nature of the judgement. This is not an intuitively easy thing to achieve within judgement practice. Accordingly, the 'flow' conceptual metaphor as presented above is one such way in which teachers and students navigate this challenge, by drawing on a familiar conceptual framework to account for what is otherwise fuzzy and indefinite. The idea of meaning making by means of comparison made possible through metaphor can perhaps account for why it has been employed by teachers to help articulate what they value in creative writing. If successful judgement practice is contingent on a tacit understanding of what makes a 'good' piece of work, we can observe that metaphor is a device through which this tacit understanding is communicated.

# Chapter 6: Conclusion

## Assessment practice

In order to understand the findings presented in Chapter Four and discussed in Chapter Five, we need to trace back to the very underpinning practice that teachers and students were undertaking in the first place: assessment. As discussed in Chapters One and Two, assessment practice is synonymous with the practice of forming a judgement, which itself is complicated and nebulous. The sometimes indistinguishable form of what assessment practice comprises can be attributed to the vast number of ways in which it can be applied in a range of contexts. Despite this we can look again to Boud's notion that all assessment practice hinges on 'the capacity to evaluate evidence...to draw sound conclusions' (2007:1). This research has sought to explore the process that teachers and students undertake when assessing and to reveal insights into the nature of an assessment decision to ultimately better understand the benefits and challenges of employing a different approach to assessment than conventional means, namely the use of comparative judgement. This final Chapter examines this enquiry's findings by revisiting the ideas central to it: assessment, standards and judgement practice. The Chapter closes with final concluding remarks and recommendations.

## Standards

This enquiry examines judgements that have been formed through comparison. One significant distinction between comparative judgement and absolute referencing conducted with reliance on mark schemes is their respective approach in interacting with standards. Absolute referencing is made possible through the presence of external standards that define and prescribe what graduated levels of performance in an activity look like. Standards in an absolute referencing model are very often decontextualized. It is for the teacher to interpret the standards, make meaning with a synthesis between their interpretation and their understanding of the subject, and then apply this in order to successfully arrive at a judgement. In formal assessment scenarios, as with GCSE English creative writing, a teacher might be required to traverse ten or more standards while undertaking the above. We can recognise this process to be a demanding one.

The relationship with standards in comparative judgement is a different one. Built into the design of many comparative judgement exercises, as is the case in this enquiry, is the omission of external standards in helping teachers form judgements. In respect to the example of absolute referencing we can perhaps take this to be advantageous, in that standards can be problematic to interpret, make sense of and apply. So, the question becomes are we to take it that no standards are present when an assessment decision is reached in a comparative judgement paradigm? At least at first glance, findings from the interviews with teachers certainly seemed to suggest this:

In response to the question: *'what helped you arrive at the decision?'*

*Teacher 8: "I enjoyed it more. It has suspense, it's structurally much more engaging than text B."*
*Teacher 9: "It's a mixture of the flow and the content, really."*

*Teacher 10: "Probably experience. I've read a lot of books and if a book doesn't interest me...doesn't pull me in...then I'm not interested. When you read something there has to be a draw...there has to be something to pull you in."*

In each of these responses we can note lack or absence of any unified standards being referenced. This is perhaps to be expected considering teachers were asked to form judgements with no reference to a mark scheme or other similar documents. But to assume that the judgements that were reached were done so without reference to any standards is, I would argue, erroneous. In each of the teacher responses above we can observe how ideas about the quality of what students had written were articulated. We can look again at Sadler's definition of the meaning of standard in an educational context as a set of 'fixed points of reference for assessing individual students' (Sadler, 1987:191) and recognise that teachers in the study clearly formed their judgements by drawing upon fixed points of reference. Examples from the quotes previously cited include "*enjoy[ment], suspense, structurally ...more engaging, flow, content, experience."*

These examples of what teachers valued are indicative of a kind of standard, albeit these have not been uniformly agreed, externalised and codified in writing as with

mark schemes. Moreover, they are not fully realised, at least in the manner articulated by teachers in these interviews. The extent to which reading an artefact of student work brings *"enjoyment",* for example, provides an insight into what one such standard might focus on. But teacher 8 provided no discriminating elements beyond reference to the word "enjoyment" to indicate how varying degrees of enjoyment might represent different levels of proficiency for them to consider when assessing. Furthermore, we can observe the challenge of attempting to quantify something personal and interpretative like levels of enjoyment. Ultimately, this is only a problem in an assessment paradigm in which absolute referencing is employed. Comparative judgement faces no such challenge as it enables teachers to consider their own internalised standards of quality and merely make a comparison judgement between two items.

The advantages posed by such comparisons are evident if we expand the quote from Teacher 8 on the importance of "*enjoy[ment]"*. Teacher 8 articulated that what helped them arrive at the decision between the better of two creative scripts was considering which one they enjoyed "*more".* The use of *"more"* here refers to the comparative nature of the judgements they were making, and this is significant. In essence, the comparative nature of the assessment design in this enquiry allowed teachers to successfully consider and draw upon internalised standards that underpin their understanding of what makes a good piece of creative writing written by students at a GCSE English level of proficiency. That these understandings are highly interpretive and individual is not problematic. We can resolve here that standards were present in the forming of teacher judgements in this enquiry, with the comparative nature of the assessment design permitting teachers to draw upon

internalised standards and tacit understandings of what we mean by good quality work in this context. These findings lend support to one of Maxwell's (2001) definitions of standards as being 'arbiters of quality (relative success or merit)', with Maxwell's assertion that standards can inform how the quality of an item is relative in view of the other items to which it is being compared against.

In consideration of the above findings, we can perhaps conclude that teachers are enabled through comparative judgement to successfully draw upon internalised quality standards when forming assessment decisions, and that the importance of externalised and codified standards is resultantly diminished. It might be argued that standards of such a construction have no place in informing what makes good creative writing. Indeed, teachers articulated in their interviews that the negotiation of standards through application of the mark schemes in an assessment task led to challenges, principally that they fostered a perception of there being an inhibitor on what they could form a judgement on:

> *Teacher 10: "In English we're assessing against a mark scheme against all the criteria, SPAG* [spelling, punctuation and grammar] *and all the rest of it, but actually sometimes when you're creative that kind of goes out the window, because you're not thinking in a uniform way.*

> *Teacher 9: "It's* [comparative judgement] *a lot better than sitting there with a mark scheme, which can drive you up the wall sometimes because you sort of know where to put a piece*

*when you look at it, and then see how the mark scheme fits around it."*

*Teacher 11: "With a mark scheme you break it down at an earlier stage, it seems as though you're compartmentalising it in a way*

These points are important in that these teachers are drawing attention to how the application of the mark scheme is a demanding and difficult task and it can carry with it inherent flaws pertaining to the reliability and validity of the judgements arrived at, as has been argued in this enquiry in Chapters One and Two. But if we are to be equitable in our challenge of examining effective assessment practices then we need to level the same critiques to comparative judgement as have been levelled at absolute referencing.

While the findings from the use of comparative judgement with teachers in this enquiry have indicated that this method is reliable, the question remains as to the validity of the judgements formed. Chapter Five tentatively presents that the high agreement values in methods 1 and 2 are representative of a shared understanding between teachers as to what good creative writing comprises, or 'what we mean by good work' (Sennett, 2008). Accordingly, it is possible to conclude that the judgements were valid. However, there is little evidence of similar agreement between teachers if we attempt to align what they spoke about in their interviews, focusing on what helped them decide the better items of work when comparatively judging. We might have expected some conformity of the ideas being shared in

interviews. What was uncovered, however, varied greatly in some instances and was largely interpretive by nature. This is seen in references to the *'enjoyment'* they felt when reading*, 'engagement'* with the text*, and 'experience'* as a basis for forming a judgement.

So, what can we take from this? Is it possible to claim that the judgements formed through comparative judgement in this enquiry are valid? This is where we can locate the benefit provided by external codified standards. With these it would be possible to determine conformity of a judge's assessment practice with the agreed standard, and if the judge's decisions, or decisions across all judges, were valid. We can recognise here the challenge that introducing a standard with which judges appraise the quality of item against would present. The effectiveness of the externalised standard on informing judges of the respective quality of an item hinges on the judges all successfully interpreting the standard in the same way. This is a very difficult thing to do. In essence in the scenario described above we have arrived back at an assessment situation similar to absolute referencing in its design and find ourselves in something of a paradox. In order for teachers to reliably judge the quality of a highly interpretive item of work such as creative writing, opportunities should be made for them to draw on their tacit understanding of what makes a good piece of work. This is as interpreting standards alone to inform judgements leads to irregularities; as Sennett (2008) notes, the risk in adhering to such standards is that there is a genuine loss of craft. But in order for the judgements that are reached as a result of comparative judgement to be deemed valid there needs to be an external reference point, through which a specific definition of quality is represented in

respect of the judgements formed. This is what I term the *defensibility / validity* paradox, in which increasing one of these leads to a reduction in the other.
This debate is centred on ideological matters that exist within creative writing as a subject in discipline of the English Language and is based on the premise that what makes good creative writing is exceedingly difficult to capture through written standards (D'Arcy, 1999; Cremin & Myhill, 2013). This has a long-standing tradition in English as a discipline. The Newbolt Report of 1921 reported that English:

> *'...connotes discovery of the world by the first and most direct way open to us, and the discovery of ourselves in our native environment...For the writing of English is essentially an art, and the effect of English literature, in education, is the effect of art upon the development of the human character'*
>
> (1921:21)

In contrast, attempts to define what 'good' and varying other levels of quality are in respect of creative writing and other facets within the study of English are common place, of which the National Curriculum represents just one. The premise in such traditions is that English comprises a determinable set of knowledge and skills, through which students make linear progress as they develop in proficiency. These two premises represent two different ends of the same ideological continuum. They also provide an interesting point for us to consider as we examine teacher and student understandings of good quality work in creative writing. We might ask which ideology dominates? We might note that a teacher or student's response to the question of what they value in creative writing would vary significantly depending on their ideological subscription towards the subject. Is it the working of a theme? A character driven narrative? The successful employment of several rich and varied

language techniques? Allegorical allusions to other works? Technical accuracy in spelling and grammar? All of these, or some combination of them, alongside a myriad of other possibilities?

These differences in how English can be understood as a subject are critical in providing us with insight into the compromises that have been made in order to avoid the *defensibility/validity* paradox. Before exploring an example, we can recognise the need for compromise; Sennett (2008) asks:

> 'what do we mean by good-quality work? One answer is how something should be done, the other is getting it to work. There is a difference between correctness and functionality. Ideally, there should be no conflict; in the real world, there is.'
> (2008:45)

In traditions such as the one seen in the National Curriculum the defensibility of a judgement has been considered paramount to that of validity. This is not necessarily a conscious decision, or an abandoning of assessment validity with forethought. Indeed, the assessment standards in the National Curriculum are predicated upon the ideological assumption that quality can be accurately assessed by their application. Rather, this reflects on the wider educational tendency to consider complex and subjective disciplines as being codifiable, and the belief that teachers should be capable and consistent interpreters of these externally set standards. The challenges identified above are largely attributable to the practical difficulties in defining and subsequently applying externally set and regulated standards. On the conflicts in measures of quality, Sennett (2008) notes:

> ‘To take a generous view, the reformers of the NHS are crafting a system that works correctly, and their impulse to reform reflects something all about craftsmanship; this is to reject muddling through, to reject the job just good enough, as an excuse for mediocrity. To take an equally generous view of the claims of practice, it encompasses pursuing a problem...in all its ramifications. This craftsman must be patient, eschewing quick fixes. Good work of this sort tends to focus on relationships; it either deploys relational thinking about objects or...attends to clues from other people. It emphasises the lessons of experience through a dialogue between tacit knowledge and explicit critique’ (2008:51).

Sennett’s reference to the NHS’s prescribed diagnostic procedure that doctors and nurses must abide by in neglect of their own knowledge and experience in order to preserve minimum acceptable standards of performance in a practice can be understood as a similar procedure to the use of assessment standards in the National Curriculum. A sympathetic view recognises the intention of these, to prevent negligence and striving beyond good enough. What is perhaps lost though, as Sennett observes, is relational and dialogic collaboration, and the utilisation of experience that is underpinned by tacit knowledge to ultimately lead to genuine craft in practice. In the context of assessment practice, this manifests as the conferring of a judgement with greater validity.

As observed in this enquiry, teachers are capable of drawing on their own internalised standards of what is meant by good quality work. These standards are their own, and have been developed through collaboration with others, through experience, through the honing of their practice. The next section of this Chapter

explores teachers' articulations of what they felt represents good quality creative writing, what we can recognise as standards of a sort, in greater depth. In doing this we take a conscious step in not seeking to resolve that the judgements formed in this enquiry are valid. Attempting to do so requires us to attempt to align judgements with external standards, the use of which can infringe on an individual's judgement practice. With this step, we are given greater liberty with which we can explore the full breadth of assessment practice.

## Judgement practice

We have determined that in this enquiry comparative judgement was successfully undertaken by teachers through the employment of tacit understanding and internalised standards of quality. We can also recognise that there is a relationship between standards across different teachers, as teacher agreement was high in view of the Rasch analysis in methods 1 and 2, but that this relationship is complex and diverging, owing to the different ways in which teachers articulated what to them represented good quality work. As such, we can identify the need to interrogate what this enquiry has revealed about judgement practice. We can look again to Dunne's (2005) definition of a practice as 'a [...] complex set of activities [...] alive in the community who are its insiders [...]. Central to any such practice are standards of excellence, themselves subject to development and redefinition (2005:152-153). Dunne asserts that practice exists as activities that are sustained through collaboration within a constantly evolving and developing community, and that standards of excellence are central to this. These standards are owned by the practitioners who inhabit and give form to the community. They manifest in the

actions of the expert practitioner. As Gregson and Todd (2019) observe, with reference to Sennett's (2008) considerations on what makes good quality work, 'these standards of practice are best passed on when they are embodied in a human being through shared practice and mutual engagement in the exercising of professional judgement in context, rather than in a lifeless, static code of practice' (2019:7). In essence, judgement practice is socially-situated and its application in any given context is dynamic, evolving, and responsive to the item to which it is being applied.

Both Sennett and Dunne observe the importance of social collaboration, cooperation and co-ownership to the development and fostering of practice. We can observe with respect to this research that there was little scope to explore how teachers might use adaptive comparative judgement over an extended period of time, and the resulting impact this would have on their practice. Significantly, there was not any provision made during the conducting of this research for the teachers to discuss their assessment practice during the act of using comparative judgement; rather, discussions took place afterwards. In this regard we can note that the findings gleaned from the ACJ trials and interviews with teachers offer us a snapshot insight into judgement practice as it was employed by teachers in specific moments in time. It is perhaps reasonable to assume that a more longitudinal exploration of comparative judgement and the impact it has on teacher judgement practice would be worthwhile.

Despite the lack of insight that a longitudinal examination of judgement practice would provide, we can recognise that there is evidence that indicates that the

teachers in this enquiry demonstrated the application of judgement practice that follows from the principles outlined by Dunne (2005) and Sennett (2008). Earlier in this chapter it was posited that the similarity in median judgement duration observed by teachers in Method 2, in which teachers undertook adaptive comparative judgement in a shared location, could be attributable to the norms dictated by the group. It could be argued that this norm came about through the co-operative and cumulative nature of the assessment design, with all teachers completing this task in the same shared environment. In an instance where teachers were faced with a new assessment paradigm they subconsciously standardised the norms for the activities and tasks the practice entailed.

Another example of Dunne's (2005) and Sennett's (2008) principles evident in judgement practice was in the approach some teachers took with regards to considering the text as a whole when undertaking comparative judgement. Teacher 11 spoke extensively on this:

> Teacher 11: *"It allowed you to take a moment and appreciate it as a piece of creative writing, rather than immediately going in for the critiquing of everything from the mark scheme"*
>
> ---
>
> Teacher 11: *"The comparative judgement gives you that opportunity at the beginning to take it all in as a whole, because it's asking "which one is better?", and that's far easier than having to tear it down to its constituent parts."*

Other teachers alluded to whole-text appreciation in referring to other holistic elements:

> Teacher 8: "*(referring to one of the texts in front of them) I enjoyed it more. It has suspense."*
>
> ---
>
> Teacher 10: *"I think you have to divide yourself - are you looking at it purely in terms of creativity? Or are you looking at it in terms of good English?"*
> Interviewer: *"which do you value more?"*
> Teacher 10: "*creativity."*

The only instruction teachers were given when comparatively judging was to select the 'most proficient text' from each combination presented to them. From this instruction came different interpretations of how quality could be best judged. Seemingly apparent in the judgement practice of these teachers was the deliberate choice to consider the quality of work presented to them in respect of the whole artefact, rather than appraising quality through the checking of constituent elements. This can be recognised as what Dunne terms a 'subversive' form of practice, in respect of the differences this form of judgement practice has when compared to how assessment of quality might be conducted through an absolute referencing assessment approach (Dunne, 2005:153). It would appear as though these teachers had confidence in their ability to determine text quality by considering the whole of

the text as a holistic piece, in spite of this being a radically different approach to the assessment of quality employed previously.

We can appreciate that this phenomenon, while subversive, is perhaps not surprising. Teachers in their interviews articulated the difficulties and limitations they have faced in assessing the quality of creative writing through criterion-based approaches. Teacher 12 made the illustrative point of citing Ernest Hemingway as a writer who might not 'fit the mould', but that is regarded as a highly competent and celebrated writer nonetheless:

> Teacher 12: *"It's as much a matter of feeling. I don't think Hemingway would pass most creative writing courses because he's too short spoken, but we agree that he's someone of quality writing."*

This is a challenge across the entire discipline of English, but most acutely in creative writing, in which the possibilities are boundless beyond the 'honesty of the writer and the scope of their imagination' (Morley, 2007:1). Accordingly, we can appreciate how some teachers focused on the whole-text, considering matters of *'feeling'* as cited by Teacher 12. This trend points to the adopting of an aesthetic perspective that viewed each text as a form of art to inform their judgement practice.

Eisner (2002) grapples with what it means to create art, noting, 'The linguistic act is the product of a linguistic imagination. The attitude required to use language of this kind is one that eludes the limiting constraints of literalism in perception and allows one to enter the work emotionally' (2002:88). For Eisner, imagining words on a page

cannot be constrained by the 'literalism', the mechanics of writing and perceiving. Rather, he advances the view that art is about 'judgement in the absence of rules. Indeed, if there are rules for making such choices, judgements would not be necessary' (ibid:77). In essence, imagination and judgement must work in tandem. To disregard an imaginative response to reading a creative writing script is to compromise the quality of the judgement. In the absence of rules, Eisner notes that 'work in the arts, unlike many other rule-governed forms of performance, always leave the door open to choice, and choice in this domain depends on a sense of rightness' (ibid:77). These concepts of 'choice' and 'a sense of rightness' are central to judgement practice in the context of art-based disciplines. For Eisner these depend on an appreciation of the aesthetic and artistry, which:

> 'Consists of having an idea worth expressing, the imaginative ability needed to conceive of how, the technical skills needed to work effectively with some material, and the sensibilities needed to make the delicate adjustments that will give the forms the moving qualities that the best of them possess' (ibid:81)

The 'material' here refers to the medium of words, in the context of this enquiry. Of significance here are references to an 'idea worth expressing', 'the imaginative ability', 'technical skills' and the need for 'delicate adjustments'. We can note that whole text appreciation promoted and made possible through ACJ enables these qualities to be considered. The 'choice' and 'sense of rightness' that a text holds hinges on the successful realisation of these by the student writer. On 'rightness', Marshall (2011) observes that 'one builds up a repertoire. Implicit within the term is a sense of a body of knowledge acquired through exposure, experimentation and

practice [...] Above all it means that English judgement is practised and criticism exercised' (2011:10). We can recognise the commonalities between 'rightness' that Marshall references, and tacit understandings of what 'good' looks like that are developed over an extended duration of continued practice.

This research recommends that teachers of GCSE English practicing in the Further Education sector need to understand and appreciate creative writing as an art-based discipline in the study of the English Language, rather than a technical set of skills or a codified set of knowledge that students must acquire. While it is recognised that some aspects of creative writing comprise technical skills and codified knowledge, these alone do not account for the full breadth of what creative writing is, or do little in informing teachers what 'good' creative writing looks and feels like. The intangible, tacit qualities that make up a 'sense of rightness' as Marshall (2011) writes makes creative writing such a unique, rewarding and celebrated pursuit. Teachers of English in the Further Education sector might contest the inclusion of creative writing as a mandatory part of the curriculum for some studying in post-compulsory education contexts. As Chapter One notes, English in the Further Education sector has often been understood through a transactional lens. Such perceptions run the risk of devaluing the ability that a command of language has to act as a means of human expression, meaning-making and self-actualisation. A further recommendation is for Further Education GCSE English teachers to develop, share and extend their tacit understandings of what 'good' creative writing looks like. This is featured below in the concluding remarks of this chapter. It is argued that collaboration, co-operation and dialogue are all valuable enablers in this process.

## Section summary

The above section of this Chapter considers what this enquiry has revealed to us about judgement practice. It has charted how teachers arrived at judgements with no reference to external standards, but through consideration of student performance in respect of their own internal standards as arbiters of what represents good quality work. It also notes the difficulty in reconciling the defensibility of a judgement with its purported validity as an assessment judgement, owing to the difficult relationship that exists between creative writing and the codified standards that attempt to define it. In addition, it has determined that judgement practice is a socially-owned and constantly developing set of tasks and activities, in which standards of what good looks like are co-constructed by members of the group. These evolve according to the responsive needs of the group and the practice itself.

Teachers in this enquiry report that they focused on whole-text quality, rather than focusing on constituent parts, when assessing through comparative judgement, and this represents one such evolution in their judgement practice. This was, it is argued, in order to evaluate the quality of creative writing in respect of the imaginative response elicited by student texts. Detaching judgement from the imagination is to suppress potential merits that the text might have. In view of judgement and imagination working in tandem, judges must draw on a 'sense of rightness' (Marshall, 2011:10). This can be understood as a form of tacit understanding.

## Concluding remarks

This enquiry set out to explore the benefits and challenges of using adaptive comparative judgement (ACJ) for the assessment of students' creative writing scripts in a Further, Adult and Vocational Education (FAVE) setting. The findings presented in this enquiry indicate that ACJ does provide benefits, including the transparency of judging decisions, the speed at which judgements are reached, and the reliability of the judgements reached. This enquiry did not set out to provide insight into how ACJ compares with traditional forms of assessment in respect of the above factors. This was in part due to the extant literature on the application of ACJ in assessment practice reporting that these benefits are replicable across different settings. The findings in respect of transparency, judgement speed and reliability are nonetheless significant and indicate that ACJ is viable and worthwhile as a mode of assessment for GCSE English creative writing scripts in a FAVE setting.

A further benefit presented by the use of ACJ for the same function is more deeply rooted in the practice of assessment. A recurring theme in this enquiry has been the prominence of tacit knowledge, the role that it plays as assessors form their judgements, and its importance in ensuring that judgements are valid. The use of ACJ as a mode of assessment has helped to uncover examples of how this tacit knowledge manifests through language, which has capably demonstrated how complex the assessment of creative writing is. The benefit in using ACJ, against what might be argued to be a challenge posed to broader assessment practices, is that valid judgements require assessors to be able to successfully call upon and apply this tacit knowledge when forming a judgement. ACJ allows a comparison

between two texts that removes assessment criterion from consideration, permitting this tacit understanding to take a more central role.

The relationship between assessment criteria and the assessment of creative writing is a tangled one. In Dunne's (2005) *Back to the Rough Ground,* he examines the concept of what he calls 'technical reason' and how appropriate it is to provide guidance for us in complex areas of life. This examination is in part a response to what he deems to be a dissatisfaction with the increasing prevalence in education to define and enact standards, focus on outcomes and increase accountability. For Dunne, the problem stems from a tendency to elevate what the Classical Greek thinkers termed *techne*, a form of scientific reason, to one of universal applicability that is capable of revealing to us all aspects of rational human action (ibid). In this fashion *techne* offers knowledge on what constitutes good quality creative writing that can be recorded through criterion, and applied to any given text in a procedural fashion in a way that leaves 'nothing to chance'. The problem is that in practice attempts at defining universal criteria encounters significant challenges.

In response to these challenges, Dunne (2005) states a distinction between *techne* and *phronesis*, a form of practical wisdom. This is characterised by 'sensitivity and attunement' towards its subject-material (ibid:256). Rather than being separable from experience, *phronesis* is realised through experiences, and is open to new experiences. Accordingly, it is made possible by negotiating with one's experience and judgement, rather than adherence to rules or criteria. It is in this conception of *phronesis* that we can recognise to be critical in underpinning and informing the process of making a judgement on the quality of a creative writing text. We can look

again at the Newbolt Report's (1921) assertion that 'English is essentially an art and the effect of English literature, in education, is the effect of art upon the development of the human character' (1921:21) as a reminder of the nature and purpose of creative writing. This must not be lost as we consider the most effective way of forming a judgement of the quality of a student's creative writing text. The 'sensitivity and attunement' that *phronesis* provides the teacher assessor is critical in an assessment scenario so their experiences, including those that are being formed as a result of participating in that specific assessment decision, can shape the judgement they are reaching.

The risk here is that detaching the assessment decision from any formal of external standards, even those that might be vaguely defined and permit some flexibility, might lead to unreliable assessment judgements. What this enquiry has found is that even when teachers form judgements without reference to external standards during the assessment process, their judgements are reliable. We can note from the findings above that the teachers in this enquiry have their own internal conceptions of 'good quality' that do share commonalities with one another. Strikingly, it has also been determined that there are shared common understandings between teachers and students in respect of 'good quality' creative writing, and that the quality markers teachers identified even bared a significant degree of similarity with those conceived of by Britton (1950) and his colleagues in research that sought to explore similar matters nearly seventy years ago. These findings tentatively point to the presence of shared understandings of what makes good quality writing, chiming with Sadler's conception of *guild knowledge.* Defining factors that makes *guild knowledge* distinct from codified assessment standards is that it is community-owned and thus open

and responsive to adaption, contextually rich as it given life through interactions between teachers and students, and exists within the practice of assessment rather than aside it in an abstracted theoretical plane.

## The importance of this research, and its original contribution to knowledge

This section considers the importance of this research and how it offers an original contribution to knowledge. In doing this, attempts are made to characterise how this research has followed from previous theoretical and empirical work, and in specific instances has taken forward specific lines of enquiry that push beyond existing works into new forms of understanding assessment practice.

As presented in Chapter Two, there exists some extant research that explores adopting adaptive comparative judgement approaches for the purposes of assessment. The thesis describes how Pollitt (2004, 2012a, 2012b) has led work in Primary and Secondary settings, across varying subjects, in this field of research. The thesis also explains how NoMoreMarking have led a considerable initiative in engaging Secondary School GCSE English teams in the use of ACJ for assessment in recent years. Other work has seen ACJ adopted in Higher Education settings across subject disciplines for different purposes (Hardy et al., 2015; Bartholomew, 2017). While these works are valuable and important, this existing research in the field of ACJ falls short of providing an enriched picture of how this approach to assessment might be implemented in the FAVE sector. This is largely due to the

unique conditions each sector of education operates within. This thesis extends conclusions drawn from research into the use of ACJ in Primary, Secondary or Higher Education settings into an FAVE context. This contribution to knowledge resides in the way in which ACJ has been implemented in a completely new sector and context. To overlook the importance of context would be to misrepresent and misunderstand how teachers' practice is fundamentally framed by the context in which they operate.

It is here that we can recognise the value of localised, context-situated research that examines and explores aspects of teacher practice that provide new insights where previous research has not yet broken ground. This enquiry has attempted to do that. It has examined and explored the use of ACJ in the assessment of GCSE English in the FAVE sector. It represents a unique mode of enquiry into a previously unresearched area. As a form of practice-focused educational research, the findings presented here need to be understood with reference to the context in which they have arisen. That is to say with an appreciation of when and where the research took place, and who participated in the study. It is argued that the findings of this enquiry provide new knowledge into the use of ACJ as a form of assessment practice, with an appreciation of the research context.

Throughout this study efforts have been made to engage the participants, both teachers and students, in the process of this research. This has included sharing preliminary findings after data was captured during the research process, and remaining in an ongoing dialogue with the teachers involved in this study regarding what they perceived to be the value, benefits and challenges of using ACJ for

assessment practice. A key purpose of this research is to ensure that it has some value for the participants involved as well as for wider communities of assessors. This is a critical feature of practice-focused research, in that it can provide experiences through which a greater understanding of one's own practice can be gleaned. Dunne's (1993) interpretation of phronesis foregrounds the importance of experience in helping to build a capacity to do the right thing at the right time for the right reasons. The nature of human experience is crucial in defining what is subsequently gained as a result of studies of experience. Heilbronn (2011) notes that 'these elements make no sense, have no meaning, bear no significance to the practitioner, until and unless they are integrated and able to be applied. Understanding develops through the practical situations in which novices are placed, and with which they grapple' (2011:7-8). This study has engaged practicing GCSE English teachers in a trial of a new mode of assessment when judging the quality of creative writing scripts. Through this engagement they 'grapple' with their own understanding of what good quality creative writing is and how we can know it when we see it. Moreover, this thesis has demonstrated that teachers' own understanding of good quality creative writing is crucial to the practice of assessment.

On the matter of assessment practice this research has charted new ground regarding what we can understand this term to mean, through examination of its relationship with creative writing, an inherently subjective field of study within the discipline of English Language. Dunne's (1993) definition of practice helps us to appreciate the complexities of what goes into forming an assessment about the quality of a creative writing script. This thesis demonstrates that tacit knowledge plays a crucial role in assessment practice when forming a judgement of quality. On

matters of quality, Sennett (2008) reminds us that adhering to highly prescribed external standards to form assessment judgements can lead us to settling for mediocrity or 'just good enough', obscuring the crafting of judgement through collaboration, cooperation and dialogue. This thesis provides a justification for all three to play a more central role in assessment practice.

On a national level the assessment practice which forms the focus of this study centres on creative writing. Conventional approaches to the assessment of creative writing remain a largely prescriptive and technical act. This is even more so in respect of summative assessment and its required and rigid adherence to assessment standards and criteria. This thesis argues that widely established and taken for granted assessment procedures are insufficient in defining good quality in creative writing. It is argued in the thesis that this reduces assessment practice to a technical act rather than an aesthetic one centred on human expression and a deep connection to the human condition. Oakeshott (1972) reminds us that education is a 'transaction between generations' (1972:63) and a deliberative activity, and the composing, rehearsing and sharing of authored stories is perhaps one of the oldest human transactional methods to exist. This thesis shows that teachers, through their assessment practice, are capable of drawing on their own tacit understandings of what good quality without standards and prescribed criteria looks like and feels like to the reader. It is argued that aesthetic engagement is a central tenet in this, and that aesthetic engagement need not detract from the reliability or validity of the judgements formed.

Further to considerations of assessment practice, this thesis has examined the concept of tacit knowledge and argues that it underpins a judgement by giving teachers opportunities to verbalise their tacit understandings of what makes good creative writing. The semi-structured interviews with teachers generated rich and revealing dialogue providing deeper insights into the processes involved in arriving at assessment judgements. It is important to note how this approach to assessment created spaces in which teachers could engage in open and honest discussions centred on their assessment practices in GCSE English. What has emerged as a key finding in this research is the value that dialogue, cooperation and collaboration of this kind can offer teachers. Traditional standardisation practices place assessment standards at the centre of the activity. In such instances teachers work individually or collaborate to align their judgement with pre-set assessment standards. What this traditional practice neglects as a result of preoccupations with written assessment standards is dialogue centred upon teachers' own understandings of what we mean by good work (Sennett, 2008) in the context of GCSE English creative writing. This aspect of assessment practice through dialogue, cooperation and collaboration is fundamental, as it is this that is drawn upon when teachers interpret prescribed assessment standards. If teachers understanding of a criterion is partial or not present then a valid interpretation of standards is not possible.

The complexities of such forms of tacit knowledge are explored in depth throughout this research. One illustration of this has been through the examination of teachers' and students' use of figurative language in discussions about creative writing script quality. The thesis employs Lakoff and Johnson's (1980) *Conceptual Metaphor Theory* in considering the implications for our understanding of tacit knowledge. It

considers what conceptual metaphor theory tells us about how we understand the world, and the manner in which tacit understanding of complex and nebulous ideas such as what me mean by good work in creative writing is communicated. Data in the study reveal how through these linguistic flourishes teachers are expressing ideas and concepts that go beyond literal definition but that are real and tangible nonetheless. The thesis discusses implications for pedagogical practice in some depth. Sadler's (1989) concept of *guild knowledge*, which comprises 'knowing ways to download evaluative knowledge to students' (1989:141) is also considered and discussed in some detail. While Sadler's use of the word 'download' may be overly mechanical the idea of knowing ways to communicate evaluative knowledge to students is vital. The challenge facing assessors is to find ways to go beyond cognitive concerns to embrace affective engagement with the reader. This takes us into new territory.

If we appreciate that an accurate tacit understanding of what good work looks like in context is an important aspect of assessment practice which all GCSE English teachers need to have, then we can note the value in using Adaptive Comparative Judgement as a method through which challenges to teacher understanding can be posed in context, as has been demonstrated in this research. Furthermore, if we appreciate that this tacit understanding is community-owned (Lave & Wenger, 1991; Sfard, 1999), and that is evolves cumulatively and cooperatively over time (Dunne, 1993), and articulated through figurative language (Lakoff & Johnson, 1980), then we can recognise the need for teachers to participate in cooperative and collaborative dialogue with their colleagues that enables the exchange of ideas in communities of assessment practice. Here Dunne (1993) is drawing our attention to how

communities of practice stay alive through the sustained commitment of their insiders, their genuine practitioners, to creatively develop and extend it, sometimes by shifts which at the time may seem dramatic or subversive. In some respects this thesis represents an attempt to do just that. It challenges taken for granted assumptions about the value of assessment practices based upon systems of criterion-referenced assessment. Dunne (ibid) goes on point out that central to any practice are standards of excellence, themselves subject to development and redefinition, which demand responsiveness from those who are or are trying to become practitioners. Once again, this thesis represents an attempt to develop and redefine standards of excellence in assessment which are capable of going beyond cognitive concerns and highly prescribed written assessment criteria.

Activities such as those described above are central to the development and sustaining of effective communities of assessment practice, which lead to teachers becoming more problem-attuned (Chinn, Maeve, and Bostwick, 1997; Sennett, 2008; Aristotle, 2011). If we are to value the defensibility of an assessment judgement, as identified as an emerging theme in Chapter One, then teachers must engage in dialogue that actively challenges and simultaneously enriches their own understanding of 'what good looks like'. This thesis indicates that such discussions are effective when facilitated in context through comparative judgement.

What is being suggested here is a slight shift in the way that GCSE English assessment standardisation is conducted and how collaboration between teachers takes place in view of perceived shortcomings in assessment practices. Chapter One chronicles a critical incident that highlights disparities in GCSE English assessment

practice that were evident in a previous mock exam window. This incident has served as a frame for this study and signals shortcomings, in the way that standardisation practices were previously ineffective in ensuring GCSE English teacher judgements were consistent and valid when assessing student performance. Examples like this illustrate the need for the trialling of alternative approaches to assessment practice. Coffield (2008) invites teachers to consider "what practices should we as teachers be holding onto and which ones should we be abandoning?". Here he is pointing to a need to ensure that educational practices are sustained only when they are genuinely educational for all those involved, rather than continuing with outmoded practices which do not represent what we mean by good work. This thesis addresses this question in the context of assessment practices. It offers a potential alternative based on the findings of this research.

## Summary of recommendations and next steps

This final section of Chapter Six features the presentation of a set of a series of recommendations that have emerged from this enquiry. These seek to define what the application of adaptive comparative judgement approaches to assessment in Further Education settings might look like, and what benefits it can potentially yield. It is important to note at this point that this small-scale study is limited in terms of generalisability. It is hoped however that the insights offered in the thesis provide the reader with a sense of the trustworthiness of this research and the authenticity of its findings. It will of course be for others in wider communities of assessment practice

to determine the extent to which the findings reported here may be of use and value elsewhere in other contexts.

The first recommendation advocates the wider adoption of adaptive comparative judgement as a mode of assessment of GCSE English creative writing scripts in the FAVE sector. The first chapter of this enquiry began with a critical incident that provided insight into the context and problem on which the subsequent exploration of ACJ as an alternative to conventional assessment was based. The intention was to explore if ACJ as a mode of assessment could provide reliable judgement decisions while at the same time not compromise the time taken per judgement. The findings of this thesis lend support to the claim that ACJ may offer a potential way to address both challenges.

In respect of specific examples as to how ACJ might be adopted within an institution we can chart some different possibilities. One function would be to mirror its implementation as has been demonstrated in this enquiry, in which teachers undertake comparative judgement on a sample of scripts, following which information about the judgements that teachers have made are reported using the NoMoreMarking software. The results that follow from this study provide a detailed insight into the judgement practice of each individual teacher, including total scripts judged, reliability in respect of other judges in the sample, and duration per judgement. These data might be useful for individual teachers, teams or leaders. From the data gathered through this method it might be possible to define specific training needs and interventions where there might not have been any before.

The value and meaning of such data, and the subsequent actions that follow from the collection and interpretation of data, will vary depending on who is reading it. As noted in Chapter Two, NoMoreMarking's claim of saving time and increasing teacher efficiency through the use of ACJ for assessment purposes is an alluring one. However applications of ACJ which are motivated by solely for the purposes of saving time may be misguided. This thesis suggests that similar benefits are possible for GCSE English teachers practicing in the FAVE sector. In addition, it argued that beyond paramount considerations of time saving and efficiency is the broader value ACJ can provide teachers in respect of the quality and value of their assessment practice. This includes the manner in which it permits the accessing and articulation of tacit understandings of 'good' quality work, how it can promote self-dialogue and critique during the process of forming a judgement, and how ACJ can promote aesthetic interpretations of students' creative writing texts. These aspects might be less immediately alluring than claims of saving time or increasing productivity, but this does not diminish their importance. Rather, it must be understood that assessment practice is complex and merits the investment of teachers' time commensurate with what is required to do the job well, including its value to learners.

So how might ACJ be effectively introduced within a FAVE setting, in view of possible conflicting motivations that differ across stakeholder groups? This thesis proposes that the ownership of ACJ practices, including the implementation and subsequent data that follow from its adoption, should lie with teachers. Of course, leaders do have a part of play here. They can create and foster the conditions in which genuine educational practices can occur. In this capacity they might establish

an environment in which ACJ can be applied in the context of their organisations. In practice this could see the use of ACJ for the summative assessment of GCSE English mock exams. However, the adoption and adaption of these practices would need to be accompanied by an appreciation that teachers own the process, and that the metrics and data that follow from the use of ACJ are not used to compare or evaluate teacher performance. One might imagine a scenario in which teachers with 'unreliable' judgements are deemed in need of some kind of formal intervention. Instead, what this research advocates is that by teachers owning the process of ACJ they are in tandem adopting an ownership and responsibility for their own professional learning. Assessment practice does not take place in a vacuum. It is defined by the context in which it takes place. Teachers are members of a community, we might argue a *guild*, and need to collaborate and learn from one another if they are to maintain and advance their practice. Ownership of the practice of ACJ is one way in which they might do that.

The second recommendation follows and builds on the first recommendation in advocating for the use of ACJ as a form of professional learning for GCSE English teachers practicing in the FAVE sector. As has been established in this thesis, the existence and importance of guild knowledge as a form of tacit understanding of what makes good quality creative writing is fundamental to good judgement in assessment practice. What has also been determined in the course of this thesis is that guild knowledge is assembled, developed and evolved in collaboration with others. It is the embodiment of experience that forms a kind of practical wisdom that cannot simply be transferred in the form an explicit instruction. So how might teachers develop this through professional learning? We can look to Sfard's (1998)

*participation metaphor* for learning here as an appropriate way of better understanding this process. The distinction between Sfard's (1999) previously discussed *acquisition metaphor* and the *participation metaphor* for learning is represented through a 'linguistic turn, the permanence of *having,* gives way to the constant flux of *doing.'* (ibid:6). In this form the learner should be viewed as a person interested in participation in activities rather than acquiring possessions, where learning is now conceived of as a process of 'becoming a member of a certain community' (ibid:6).

In view of the participation metaphor of learning, we can note the value in creating opportunities through which purposeful dialogue can occur within a community that challenge and extend its members' understandings of that community and the world. As has been determined in this enquiry ACJ can serve as a medium through which teachers are presented with challenges to their understanding of what good quality creative writing is. All teachers in this enquiry capably articulated the thought processes that underpinned their judgement decisions, which in turn were arrived at without reference to assessment criteria. Many teachers also presented evidence of engaging in a self-dialogue during the process of judgement. This evidence points to the potential value of creating spaces for dialogue between teachers in which they might collaboratively undertake adaptive comparative judgements, narrate their judgement decisions to colleagues and tackle the challenges posed by such decisions with others. Sennett's (2008) notion of becoming 'problem-attuned' through problem finding, problem solving and critique is helpful here. Such practice is not entirely different from standardisation activities that might be undertaken by a teaching team attempting to align texts to a set of standards. The difference in what I

am advocating is that this practice would be undertaken through ACJ and without reference to assessment standards, with the primary intention being to draw out and develop a mutually-owned guild knowledge.

The third recommendation advocates using ACJ with students to peer assess each other's creative writing texts. Peer assessment is a long-established tradition in classrooms that is used by teachers across multiple subject disciplines. In GCSE English classrooms, peer assessment is typically accompanied by assessment criteria which students are to use to inform their assessment decisions. Such practices can be problematic. As Polanyi (1964) observes:

> *'Maxims are rules, the correct application of which is part of the art which they govern. The true maxims of golfing or of poetry increase our insight into golfing or poetry and may even give valuable guidance to golfers and poets; but these maxims would instantly condemn themselves to absurdity if they tried to replace the golfer's skill or the poet's art. Maxims cannot be understood, still less applied by anyone not already possessing a good practical knowledge of the art. They derive their interest from our appreciation of the art and cannot themselves either replace or establish that appreciation.*
>
> *(1964:31)*

For Polanyi, the problem is that assessment criteria can become a sort of misrepresentation of what is actually happening in a piece of creative writing, in which they 'condemn themselves to absurdity' by attempting to account for something inherently subjective. Moreover, this problem is exacerbated if the assessor does not have a solid foundation of understanding of what it is they are

assessing, something commonly attributable to students, who are still apprenticing in the subject. As Coe (2019) argues, criteria ‘are not meaningful unless you know what they already mean’.

It is in this climate that we can locate ACJ as a viable and useful form of peer assessment. As ACJ facilitates comparative assessment judgements without the need for assessment criteria by design, the problem of students having to traverse the rocky path of assessment criteria identified by Polanyi and Coe is avoided. So how can ACJ be used to promote valuable peer assessment, and by extension productive learning opportunities? An important point to note here is that ACJ as peer assessment should not be positioned as a form of discovery learning, in which students undertake judgements of quality in isolation and implicitly develop expertise in recognising good quality work. Much in keeping with the principles of using ACJ advocated in the second recommendation above, it could be used for peer assessment within an open, dialogic and culturally-rich environment in which students are encouraged to articulate the judgements they are forming. A key requirement in the design of this peer assessment environment is the presence of a teacher advanced in assessment practice, an expert who can support, clarify and moderate. The goal here is to apprentice students into the *guild*, and support them in learning what good creative writing looks and feels like by modelling examples, always experienced in context.

# Bibliography

Aristotle (2011). *Nicomachean Ethics* (R. C. Bartlett & S. D. Collins, Trans). Chicago: The University of Chicago Press.

Aristotle, Poetics, 1457. *References are to De Poetica*, trans. I. Bywater, in Aristotle. Works. ed. W.D.Ross, Vol. XI (Oxford: Clarendon Press, 1924).

Armstrong, F.; Moore, M. (2004). *Action Research for Inclusive Education: Changing Places, Changing Practices, Changing Minds.* London: Routledge.

Barnes, D.; Shemilt, D. (1974). Transmission and Interpretation. *Education Review*, 26(3), 213-228.

Bartholomew, S. (2017). Assessing open-ended design problems. *Technology and Engineering Teacher*, 76(3), 13-17.

Bassey, M. (1998) Fuzzy Generalisations: an approach to building educational theory. In: *British Educational Research Association Annual Conference*. The Queen's University of Belfast, Northern Ireland, 27th – 30th August 1998.

Bathmaker-Ann-Marie (2013) Defining 'knowledge' in vocational education qualifications in England: an analysis of key stakeholders, and their constructions of knowledge, purposes and content. *Journal of Vocational Education and Training,* 65(1), 87-107.

Bell, B.; Cowie, B. (2001). *Formative assessment and science education*. Dordrecht, The Netherlands: Kluwer Academic Press.

Bell, N; Cowie, B. (2006). *Formative Assessment and Science Education*. Berlin: Springer Science and Business Media.

Berliner, D. C. (2002, November). Educational research: The hardest science of all. *Educational Researcher*, 31(8), 18-20.

Bewley, S. Smardon, D. (2007) ‘How can dialogue create opportunity for students to think and express their ideas?’. Paper presented at the British Educational Research Association Annual Conference, Institute of Education, University of London.

Black, P.; Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80(2), 139-148.

Boud, D. (2007). Reframing assessment as if learning was important. In Boud, D. & Falchikov, N. (Eds.) *Rethinking Assessment for Higher Education: Learning for the Longer Term*. London: Routledge, 14-25.

Boud, D.; Falchikov, N. (2007). Developing assessment for informing judgement. In Boud, D. & Falchikov, N. (Eds.) *Rethinking Assessment for Higher Education: Learning for the Longer Term*. London: Routledge, 181-197.

Boyatzis, R. (1998). *Transforming qualitative information: Thematic analysis and code development*. Thousand Oaks, CA: Sage.

Bramley, T. 2015, *Investigating the reliability of Adaptive Comparative Judgment.* [pdf] Cambridge: Cambridge Assessment Research Report. Available at: <https://www.cambridgeassessment.org.uk/Images/232694-investigating-the-reliability-of-adaptive-comparative-judgment.pdf> [accessed 5 June 2019]

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(1), 77–101.

Britton, J, (1950). *Report on the Meaning and Marking of Imaginative Compositions.* London: LATE.

Britton, J. (1964). *The Multiple Marking of Compositions*. London: HMSO.

Broad, J. (2015). So many worlds, so much to do: Identifying barriers to engagement with continued professional development for teachers in the further education and training sector. *London Review of Education*, 13(1), 16-30.

Broad, J. (2016). Vocational knowledge in motion: rethinking vocational knowledge through vocational teachers' professional development. J*ournal of Vocational Education & Training*, 68(2), 143-160.

Bruner, J. (1986). *Actual Minds, Possible Worlds*. Cambridge, MA: Harvard University Press.

Bryman, A. (2004). *Social research methods. (2nd ed.).* Oxford: Oxford University Press.

Carr, W. (1995). *For Education: Towards Critical Educational Inquiry.* Buckingham: Open University Press.

Chinn, P., Maeve, M., & Bostwick, C. (1997). Aesthetic inquiry and the art of nursing. *Scholarly Inquiry for Nursing Practice*, 11(2), 83– 96.

Coe, R. (2019). *Assessment (without levels) Part 2.* [video online] Available at: <https://vimeo.com/125186094> [Accessed 18 November 2019].

Coe, R.; Waring, M.; Hedges, L.; Arthur, J. (2017). *Research Methods and Methodologies in Education. (2nd ed.)*. London: Sage Publishing.

Coffield, F. (2004) *Learning styles and pedagogy in post-16 learning: a systematic and critical review.* London: Learning and Skills Research Centre.

Coffield, F. (2008). *Just suppose teaching and learning became the first priority.* London: Learning and Skills Network.

Cohen, L.; Manion, L.; Morrison, K. (2007). *Research methods in education 6th edition*. London: Routledge.

Cremin, T. Myhill, D. (2013) *Writing Voices: Creating Communities of Writers*. London: Routledge.

Creswell, J. (2012). *Qualitative Inquiry and Research Design: Choosing Among Five Approaches.* London: SAGE Publications.

Crooks, T. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, 58(4), 438-481.

Crotty, M. (1998). *The foundations of social research*. London: Sage.

Crowley, S. (2010). *Teaching Skills for Dummies.* New Jersey: John Wiley & Sons.

D'Arcy, P. (1999). *Two Contrasting Paradigms for the Teaching and Assessment of Writing.* Loughborough: Quorn Selective Repro Ltd.

Denscombe, M. (1998) *The Good Research Guide.* Buckingham. Open University Press

Denscombe, M. (2010). *Ground rules for social research (2nd ed.)*. Maidenhead: Open University Press.

Department for Education, 2016. *Educational Excellence Everywhere.* [pdf] available at: <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/508447/Educational_Excellence_Everywhere.pdf> [Accessed 24 June 2016].

Department for Education, 2014. *Maths and English provision in post-16 education.* [online] Available at: <https://www.gov.uk/government/speeches/maths-and-english-provision-in-post-16-education> [accessed 20 May 2018]

Department for Education, 2014. *National curriculum in England: framework for key stages 1 to 4*. London: Department for Education

Departmental Committee of the Board of Education [The Newbolt Report] (1921) *The Teaching of English in England: Being the Report of the Departmental Committee Appointed by the President of the Board of Education to Inquire into the Position of English in the Educational System of England,* London: HMSO.

Dorling, D. (2015). *Injustice: Why Social Inequality Still Persists*. Bristol: Policy Press.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies.* Oxford: Oxford University Press.

Duckworth, V., Brzeski, A. (2015) 'Literacy, Learning And Identity: Challenging The NeoLiberal Agenda Through Literacies, Everyday Practices And Empowerment. *Research in Post-Compulsory Education*, 20(1), 1-16

Dunne, J. (1993). *Back to the Rough Ground: Practical Judgement and the Lure of Technique.* Indiana: Notre Dame Press.

Dunne, J. (2005) '*What's the Good of Education?'* In, Wilfred Carr (Ed.) Philosophy of Education. Abingdon: Routledge Falmer, 145-158.

Education and Training Foundation, 2018. *ETF launches Maths and English Shaping Success Campaign*. [online] Available at: <https://set.et-foundation.co.uk/news-events/blogs-and-articles/set-news/etf-launches-maths-and-english-shaping-success-campaign/#> [accessed 2 June 2018]

Education Endowment Foundation, 2019. [online]. Available at: <https://educationendowmentfoundation.org.uk/> Accessed 4 April 2019

Eisner, E. (2002). *The Arts and the Creation of Mind.* New Haven: Yale University Press.

Elliot, J; Stemler, S; Stenberg, R; Grigorenko, E; Hoffman, N. (2011) 'The socially skilled teacher and the development of tacit knowledge', *British Education Research Journal*. 37(1), 81-103.

Ericsson, K.; Krampe, R.; Tesch-Romer, C. (1993) The Role of Deliberate Practice in the Acquisition of Expert Performance. *Psychological Review*, 100(3), 363-406.

ETF (2015). *1000 teachers enrol on English Enhancement Programme.* ETF news article, 23 January, <https://www.et-foundation.co.uk/news/1000-teachers-enrol-english-enhancement-programme/>, accessed 10 May 2018.

Falchikov, N. and Goldfinch, J. (2000) Student Peer Assessment in Higher Education: A MetaAnalysis Comparing Peer and Teacher Marks. *Review of Educational Research*, 70(3), 287–322.

FEWeek, 2017. *English and Maths GCSE resit results 2017.* [online] Available at: <https://feweek.co.uk/2017/08/24/english-and-maths-gcse-resit-results-2017/> [accessed 22 May 2018]

Filer, A. (2002). *Assessment: Social Practice and Social Product*. London: Routledge.

Fox, S. (2000). Communities of practice, Foucault and Actor-network theory. *Journal of Management Studies*, 36(6), 853-866.

Freire, P. (1973) *Education For Critical Consciousness.* New York: Continuum International Publishing Group.

Frowe, I. (2001). Language and educational research. *Journal of Philosophy and Education*, 35(2), 175-186.

Fuller, A.; Unwin, L. (2011). Vocational education and training in the spotlight: back to the future for the UK's Coalition Government?, *London Review of Education*, 9(2), 191-204.

Gadamer, H.G. (1975). Hermeneutics and Social Science. *Cultural Hermeneutics*, 2(4), 307–316.

Gee, J.P. (1996). *Social Linguistics And Literacies: Ideology In A Discourse (2nd ed.)*. London: Taylor and Francis.

Giddens, A. (1982). *Contemporary Social Theory*. London: The Macmillan Press LTD.

Gillard D, 2011. *Education in England: a brief history* [online] available at: <www.educationengland.org.uk/history> Accessed 17 January 2019.

Graham Maxwell, 2001. *Are core learning outcomes 'standards'?*. [online] Available at: <https://www.qcaa.qld.edu.au/downloads/publications/research_qscc_assess_report_1.pdf> [accessed 28 May 2018]

Greenhalgh, J.; Flynn, R.; Long, A.; Tyson, S. (2008). Tacit and encoded knowledge in the use of standardised outcome measures in multidisciplinary team decision making: A case study of in-patient neurorehabilitation. *Social Science & Medicine,* 67(1), 183-94.

Gregson, M.; Todd, B. (2019) *Realizing Standards of Quality in Vocational Education and Training'.* In: Handbook of Vocational Education and Training for the Changing World of Work. Springer, Switzerland.

Grigorenko, E; Sternberg, R; Strauss, S. (2006) Practical intelligence and elementary school teacher effectiveness in the United States and Israel: measuring the predictive power of tacit knowledge, *Thinking Skills and Creativity*, 1, 14–33.

Grix, J. (2002) *Introducing Students to The Generic Terminology of Social Research*. Politics, 22(3), 175-186.

Grix, J. (2010). *Foundations of Research*. London: Palgrave Macmillan.

Hardy, J.; Galloway, R.; Rhind, S.; McBride, K., Hughes, K.; Donnelly, R., 2015. Ask, answer, assess: Peer learning from student-generated content. *HE Academy.* [online] <https://www.heacademy.ac.uk/system/files/ask_answer_assess.pdf> Accessed 20 May 2018.

Hastings, S., 2003. Questioning. *TES News*, [online] 4 June. Available at: <https://www.tes.com/news/questioning> [accessed 2 May 2018]

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

Heilbronn, R. (2011). 'Practical Judgement and Evidence-Informed Practice'. In R. Heilbronn and J. Yandell (eds) *Critical Practice in Teacher Education: A Study of Professional Learning*. London: University of London, Institute of Education (IOE).

*House of Commons,* 2009. *National Curriculum Report of Session 2008-09, volume I:9.* [online] available at: <https://publications.parliament.uk/pa/cm200809/cmselect/cmchilsch/344/34402.htm>. Accessed 20 January 2019.

Hoy, A., Hoy, W. (2013). *Instructional leadership: A research-based guide to learning in schools* (4th ed.). Boston: Allyn and Bacon.

Ish-Horowicz, S. (2015). *Increasing confidence and competence for teaching Controlled Assessments on the GCSE English re-take course*. emCETT Practitioner-Led Research Programme Report, https://www.excellencegateway.org.uk/content/etf2639, accessed 10 May 2018.

Johnson, M. (1981). *Philosophical Perspectives on Metaphor*. Minnesota: University of Minnesota Press.

Joughin, G. (2008). *Assessment, Learning and Judgement in Higher Education*. Netherlands: Springer.

Kimbell, R.; Wheeler, T.; Stables, K.; Shepard, T.; Martin, F.; Davies, D.; Pollitt, A.; Whitehouse, G., 2009. *E-scape portfolio assessment phase 3 report.* [online] available at: <https://www.teachertoolkit.co.uk/wp-content/uploads/2014/08/e-scape_phase3_report.pdf>. Accessed 2 January 2018.

Kincheloe, J. (2012). *Teachers as Researchers (Classic Edition): Qualitative Inquiry as a Path to Empowerment.* London: Routledge

King, N. (2004). Using templates in the thematic analysis of text. In C.Cassell & G. Symon (Eds.), *Essential guide to qualitative methods in organizational research* (pp. 257–270). London, UK: Sage.

Klenowski, V.; Wyatt-Smith, C. (2010). Standards, teacher judgement and moderation in contexts of national curriculum and assessment reform. *Assessment Matters*, 2(1), 107-131.

Klenowski, V.; Wyatt-Smith, C. (2013). *Assessment for Education: Standards, Judgement and Moderation*. London: Sage Publishing.

Kluger, A.; DeNisi, A. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychology Bulletin*, 119(2), 254-284.

Knight, P., 2007. The assessment of 'wicked' competencies. Report to the Practice-based Professional Learning Centre [PDF] The Open University. Available at: <http://www.open.ac.uk/opencetl/sites/www.open.ac.uk.opencetl/files/files/ecms/web-content/knight-and-page-(2007)-The-assessment-of-wicked-competences.pdf> [Accessed 15th July 2019]

Kuhn, T. (1970). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lakoff, G.; Johnson, M. (1980). *Metaphors We Live By*. London: The University of Chicago Press.

Laming, D. (2011) *Human Judgement. The eye of the beholder*. 1st ed. Andover: Cengage Learning.

Lave, J., and Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge, Cambridge University Press.

Lea, M. (2004) ‘Academic Literacies: A Pedagogy For Course Design’. *Studies In Higher Education*, 29(6), 739-756.

Liamputtong, P. (2009). *Qualitative research methods (3rd ed.).* Melbourne: Oxford University Press.

Lincoln, Y., & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage.

Marshall, B. (2011) *Testing English: Formative and Summative Approaches to English Assessment.* London: Continuum.

Marshall, B.; Wiliam, D. (2006). *English Inside the Black Box.* London: NFER Nelson.

McNiff, J.; Lomax, P. (2004). *You and Your Action Research Project*. London: Routledge.

Morley, D. (2007). *The Cambridge Introduction to Creative Writing*. Cambridge: Cambridge University Press.

Morrison, K. (1994). *Implementing Cross-Curricular Themes.* London: Routledge.

Morse, J., & Richards, L. (2002). Coding. In J. Morse & Richards (Eds.), *Read me first for a user’s guide to qualitative methods* (pp. 111–128). Thousand Oaks, CA: Sage.

National Adult Literacy Database (n.d.) Background Information On The International Adult Literacy Survey (IALS). https://www.oecd.org/edu/skills-beyond-school/41529765.pdf [accessed 14 December 2017].

Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist*, 22(2), 155-175.

Networking The Networks, 2019. [online]. Available at: <https://networkingthenetworks.com/> Accessed 7 April 2019

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: How high stakes testing corrupts America's schools*. Cambridge, Massachusetts: Harvard Education Press.

NoMoreMarking. (2017). *Reliability and English Mocks.* [online] <https://blog.nomoremarking.com/reliability-and-english-mocks-3a9b5666d682> Accessed 2 June 2018.

NoMoreMarking. (2018). Judging GCSE English: Efficiency and Reliability. [online] <https://blog.nomoremarking.com/judging-gcse-english-efficiency-and-reliability-9a8df9b80096> Accessed 2 June 2018.

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic Analysis: Striving to Meet the Trustworthiness Criteria. *International Journal of Qualitative Methods*, 16(1), 1-13.

Nyquist, J. (2003). *The benefits of reconstructing feedback as a larger system of formative assessment: a meta-analysis.* Unpublished master of science thesis, Vanderbilt University, Nashville, TN.

Oakeshott, M (1972). 'Education: the Engagement and its Frustration' in Fuller, T (ed) (1989) *The Voice of Liberal Learning*. New Haven and London: Yale University Press.

Odell, L. (1993). *Theory and Practice in the Teaching of Writing: Rethinking the Discipline.* Illinois: Southern Illinois University Press.

OECD (2012). Programme For International Student Assessment (PISA): Results From PISA. https://www.oecd.org/unitedkingdom/PISA-2012-results-UK.pdf [accessed 16 December 2018].

Ofqual (2012). *Criteria For Functional Skills Qualifications*. Coventry: Crown.

Patton, M. Q. (2002). *Qualitative research and evaluation methods.* Thousand Oaks, CA: Sage.

Polanyi, M. (1958). *Personal Knowledge: Towards a Post-Critical Philosophy.* Chicago: University of Chicago Press.

Polanyi, M. (1966). *The Tacit Dimension*. Chicago: University of Chicago Press.

Pollitt, A. (2004). *Let's stop marking exams.* http://www.cambridgeassessment.org.uk/images/109719-let-s-stop-markingexams.pdf, accessed 12 May 2018.

Pollitt, A. (2012a). Comparative judgement for assessment. *International Journal of Technology and Design Educatio*n, 22(2), 157-170.

Pollitt, A. (2012b) The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice*. 19(3), 281-300.

Pollitt, A., 2015. On 'Reliability' Bias in ACJ: Valid simulation of Adaptive Comparative Judgement. [online] available at: <https://www.researchgate.net/publication/283318012_On_'Reliability'_bias_in_ACJ > Accessed 8 Janaury 2018.

Roberts, P. and Smith, M. (2014). *Make them laugh, make them cry; re-imagining the initial assessment process for GCSE English students in the Further Adult and Vocational Education sector in England* [pdf] Education and Training Foundation Excellence Gateway. Available at: <https://www.excellencegateway.org.uk/content/etf2596> [Accessed 2 September 2018]

Rust, C., Price, M., and O'Donovan, B. (2003) Improving Students' Learning by Developing their Understanding of Assessment Criteria and Processes. *Assessment and Evaluation in Higher Education*, 28(2), 147–64.

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education*, 13(2), 191-209.

Sadler, D.R. (1989) Formative Assessment and the Design of Instructional Systems. *Instructional Science*, 18(2), 119–44.

Sandelowski, M. (2004). Using qualitative research. *Qualitative Health Research*, 14(1), 1366–1386.

Scotland, J. (2012). Exploring the Philosophical Underpinnings of Research: Relating Ontology and Epistemology to the Methodology and Methods of the Scientific, Interpretive, and Critical Research Paradigms. *English Language Teaching*, 5(9), 9-16.

Scott, D., Usher, R. (2002). *Understanding Educational Research*. London: Routledge.

Scriven, M. (1970). Explanations, predictions, and laws. In B. A. Brody (Ed), *Readings in the philosophy of science* (p. 88-104). NJ: Prentice Hall.

Sennett, R. (2008). *The Craftsman*. London: Penguin Group.

Sennett, R., 2016. Richard Sennett: Craftsmanship. [video online] Available at: <https://www.youtube.com/watch?v=nIq4w9brxTk> accessed 4 October 2018.

Sfard, A. (1998). On Two Metaphors for Learning and the Dangers of Choosing Just One. *Educational Researcher*. 27(2), 4-13.

Stables, A. (1996). *Subjects of Choice: the process and management of pupil and student choice*. New York: Cassell.

Sternberg, R. (1997) *Successful intelligence: how practical and creative intelligence determines success in life.* New York: Plume.

Sternberg, R; Hedlund, J. (2002) Practical intelligence, and work psychology, *Human Performance*, 15(1/2), 143–160.

Sternberg, R; Horvath, J. (1995) A prototype view of expert teaching, *Educational Researcher*, 24(6), 9–17.

Thurstone, L. (1927). A law of comparative judgement. *Psychological Review*, 34(1), 273-286.

Tobin, K. (2006). *Doing Educational Research*. Boston: Sense Publishers.

Stobart, G. (2008). *Testing times: The uses and abuses of assessment.* London: Routledge.

Tobin, G. A., & Begley, C. M. (2004). Methodological rigour within a qualitative framework. *Journal of Advanced Nursing*, 48(1), 388–396.

Trochim, W. (2000). *The Research Methods Knowledge Base, 2nd Edition*. Cincinnati: Atomic Dog Publishing.

Uluman, M.; Dogan, C. (2016). Comparison of Factor Score Computation Methods In Factor Analysis. *Australian Journal of Basic and Applied Sciences*, 10(18), 143-151.

Vygotsky, L. S. (1978). *Mind in Society: the Development of Higher Psychological Processes.* Cambridge, MA: Harvard University Press.

Whitehouse, C.; Pollitt, A., 2012. Using Adaptive Comparative Judgement to Obtain a Highly Reliable Rank Order in Summative Assessment. [online] available at: <https://research.aqa.org.uk/sites/default/files/pdf_upload/CERP_RP_CW_20062012_2.pdf> Accessed 5 January 2018.

Wiles, R. Crow, G. Heath, S, Charles, V. (2006) 'Anonymity and Confidentiality'. Paper presented at the ESRC Research Methods Festival, University of Oxford.

Wiliam, D. (1994). Assessing authentic tasks: Alternatives to mark-schemes. *Nordic Studies in Mathematics Education*, 2(1), 48-68.

Wiliam, D. (1996). Standards in Education: A matter of trust. *The Curriculum Journal*, 7(3), 293-306.

Wiliam, D. (1998). The Validity of Teachers' Assessments. Paper presented at the 22nd annual conference of the International Group for the Psychology of Mathematics Education, Stellenbosch, South Africa.

Wiliam, D. (2006). Formative Assessment: Getting the Focus Right. *Educational Assessment*, 11(3-4), 283-289.

Wiliam, D., 2016. 10 Misconceptions about Comparative Judgement. [LearningSpyBlog, comment on article] 9 July 2016. Available at:<https://learningspy.co.uk/assessment/not-even-wrong/> [Accessed 11 November 2018]

Wiliam, D., 2019. Teaching not a research-based profession. [online] TES. Available at: <https://www.tes.com/news/dylan-wiliam-teaching-not-research-based-profession> Accessed 25 May 2019

Winch, C. (2010). *Dimensions of Expertise: A Conceptual Exploration of Vocational Knowledge*. London: Continuum.

Wolf, A. (2011). *Review of Vocational Education – The Wolf Report*. London: Department for Education.

Zimmermann, J., 2016. *Hermeneutics: A Very Short Introduction* | Jens

Zimmermann. [video online] Available at: <https://www.youtube.com/watch?v=6wPTV5hyB0Y> [Accessed 23 June 2019].

Zolla, M; Winter, S. (2002). Deliberate Learning and the Evolution of Dynamic Capabilities. *Organisation Science*, 13(3), 339-381.

# Appendices

## 8.1 Information sheet for prospective participants

### Title

**A research study led by Michael Smith, October 2017 – June 2019**

### Information for participants January 2018

I am conducting this small-scale research project as part of my Educational Doctorate of Philosophy at Sunderland University.

The investigation seeks to explore the use of Adaptive Comparative Judgement for assessment practice in GCSE English Language. The main aims are to consider:

1. *What are the benefits and challenges of using adaptive comparative judgement approaches when assessing GCSE English creative writing scripts in a Further Education institution?*
2. *What new knowledge can be acquired by teachers as a result of undertaking adaptive comparative judgement and what function does this serve teachers of GCSE English in an FE context?*
3. *How adaptive comparative judgement can be used across a team of teachers to standardise assessment practices?*
4. *What can learners' adaptive comparative judgement decisions tell us about their understanding of creative writing as a field of study in the discipline of English Language, and what are the subsequent pedagogical implications that follow from this?*

The enquiry will use three main forms of data collection: findings gathered during workshops through use of NoMoreMarking comparative judgement software, semi-structured interviews with teachers, and interviews and focus groups with students. Workshops are expected to last no longer than an hour, and interviews no longer than 30 minutes. Interviews will be recorded and transcribed.

The research will be conducted under the guidelines of the British Educational Research Association. Data will be kept confidential according to these guidelines, and participants, unless they choose otherwise, will be unidentifiable in any publications resulting from the study. Participants will be able to withdraw from the study at any time.

If you are interested in participating, or in learning more about the project, please contact Michael Smith by email as follows: Michael.Smith@bdc.ac.uk

## 8.2 Consent form for participants

### Title

**A research study led by Michael Smith, October 2017 – June 2019**

## Participant Consent form

I have been fully informed about the aims and purposes of the project.

I understand that:

- this project seeks to explore assessment practice so as to inform potential improvements;
- there is no compulsion for me to participate in this research and, if I do choose to participate, I may at any stage withdraw my participation;
- any information which I give will be used solely for the purposes of this research project, which may include publications;
- the information which I give may be reported on in anonymised form;
- all information which I give will be treated as confidential, and pseudonyms will be used in order to preserve anonymity to the greatest possible extent

| ..... | ..... | ……..... |
|---|---|---|
| (Signature) | (Printed name) | (Date) |

One copy of this form will be kept by me; a second copy will be kept by the researcher.

# 8.3 Mock performance vs. final grade in 127 students in the 2016-17 academic year, full table.

| STUDENT ID (omitted) | Main course title | Register/Course Title | English mock (paper 1) Section A = Reading, Section B = Writing | GCSE English grade obtained |
|---|---|---|---|---|
| Omitted | BTEC EXTENDED DIPLOMA IN CONSTRUCTION | GCSE English | 3 | 4 |
| Omitted | AQA GCSE ENGLISH (DAY) | GCSE English | 2 | FL - fail |
| Omitted | BTEC LEVEL 2 DIPLOMA IN FASHION DESIGN | GCSE English | 3 | 4 |
| Omitted | BTEC LEVEL 2 FIRST CERTIFICATE IN BUSINESS (TSA) | GCSE English | 4 | 3 |
| Omitted | AQA GCSE ENGLISH | GCSE English | 1 | 3 |
| Omitted | AQA GCSE ENGLISH | GCSE English | 3 | 2 |
| Omitted | BTEC LEVEL 2 DIPLOMA IN ART AND DESIGN | GCSE English | 2 | 2 |
| Omitted | CITY & GUILDS DIPLOMA FOR LEGAL SECRETARIES (LEGAL 3) | GCSE English | 2 | 4 |
| Omitted | AQA GCSE ENGLISH (DAY) | GCSE English | 5 | 5 |
| Omitted | CG 7202 LEVEL 1 DIPLOMA IN ELECTRICAL INSTALLATIONS | GCSE English | 4 | 3 |
| Omitted | LEVEL 3 DIPLOMA IN CHILDCARE AND EDUCATION | GCSE English | 7 | 5 |
| Omitted | EXTENDED DIPLOMA IN SPORT (RUGBY ACADEMY) | GCSE English | 1 | 2 |
| Omitted | BTEC LEVEL 2 BTEC FIRST CERTIFICATE IN BUSINESS (RG) | GCSE English | 1 | 4 |
| Omitted | BTEC SUBSIDIARY DIPLOMA IN PERFORMING ARTS (ACTING) | GCSE English | 3 | 3 |
| Omitted | BTEC EXTENDED DIPLOMA IN FASHION DESIGN | GCSE English | 2 | 2 |
| Omitted | FOUNDATION DEGREE IN 3D DESIGN | GCSE English | 3 | 3 |
| Omitted | AQA GCSE ENGLISH (DAY) | GCSE English | 3 | 6 |
| Omitted | LEVEL 3 BTEC SUBSIDIARY DIPLOMA IN BUSINESS | GCSE English | X | FL - fail |
| Omitted | AQA GCSE ENGLISH (DAY) | GCSE English | X | FL - fail |
| Omitted | CITY & GUILDS DIPLOMA FOR LEGAL SECRETARIES (LEVEL 2) (LEGAL 1) | GCSE English | 2 | 3 |
| Omitted | BTEC LEVEL 2 BTEC FIRST CERTIFICATE IN BUSINESS (RG) | GCSE English | 1 | 3 |
| Omitted | LEVEL 3 OCR DIPLOMA IN ADMINISTRATION (BUSINESS PROFESSIONAL) | GCSE English | 4 | 3 |
| Omitted | CG 7202 LEVEL 1 DIPLOMA IN ELECTRICAL INSTALLATIONS | GCSE English | X | 1 |
| Omitted | BTEC LEVEL 3 SUBSIDIARY DIPLOMA IN IT - SOFTWARE DEVELOPMENT | GCSE English | 3 | 2 |
| Omitted | AQA GCSE ENGLISH | GCSE English | 2 | 3 |
| Omitted | LEVEL 2 DIPLOMA IN BEAUTY THERAPY (FULL TIME) TSA | GCSE English | 2 | 3 |

| | | | | |
|---|---|---|---|---|
| Omitted | OCR LEVEL 2 TECHNICAL DIPLOMA IN SPORT | GCSE English | 2 | 3 |
| Omitted | CACHE LEVEL 2 EXTENDED DIPLOMA IN HEALTH AND SOCIAL CARE | GCSE English | 3 | 3 |
| Omitted | LEVEL 2 DIPLOMA IN BEAUTY THERAPY (FULL TIME) | GCSE English | 1 | 1 |
| Omitted | BTEC EXTENDED DIPLOMA IN CONSTRUCTION | GCSE English | 3 | 3 |
| Omitted | BTEC SUBSIDIARY DIPLOMA IN CONSTRUCTION | GCSE English | 3 | 2 |
| Omitted | OCR LEVEL 2 DIPLOMA IN IT | GCSE English | 3 | 2 |
| Omitted | AQA GCSE ENGLISH (DAY) | GCSE English | 6 | 4 |
| Omitted | AQA GCSE ENGLISH | GCSE English | 2 | 3 |
| Omitted | C&G 6035 LEVEL 2 DIPLOMA IN PLUMBING STUDIES | GCSE English | 3 | FL - fail |
| Omitted | C&G 7202 LEVEL 1 DIPLOMA IN PLUMBING STUDIES | GCSE English | 4 | 4 |
| Omitted | BTEC SUBSIDIARY DIPLOMA IN GAMES, ART AND ANIMATION | GCSE English | 4 | 4 |
| Omitted | ACCESS TO H.E DIPLOMA (TEACHING) | GCSE English | 8 | 6 |
| Omitted | EXTENDED DIPLOMA IN SPORT (RUGBY ACADEMY) | GCSE English | 2 | 3 |
| Omitted | BTEC SUBSIDIARY DIPLOMA IN CONSTRUCTION | GCSE English | 2 | 4 |
| Omitted | BTEC SUBSIDIARY DIPLOMA IN CONSTRUCTION | GCSE English | 2 | 3 |
| Omitted | NCFE LEVEL 2 DIPLOMA FOR ENTRY TO THE UNIFORMED SERVICES | GCSE English | X | 5 |
| Omitted | AQA GCSE ENGLISH | GCSE English | X | 4 |
| Omitted | C&G 2365 DIPLOMA IN ELECTRICAL INSTALLATIONS LEVEL 2 | GCSE English | X | FL - fail |
| Omitted | LEVEL 3 SUBSIDIARY DIPLOMA IN SPORT - RUGBY ACADEMY | GCSE English | 3 | 3 |
| Omitted | AQA GCSE ENGLISH | GCSE English | X | FL - fail |
| Omitted | BTEC EXTENDED DIPLOMA IN MUSIC | GCSE English | 1 | 2 |
| Omitted | LEVEL 3 DIPLOMA IN CHILDCARE AND EDUCATION | GCSE English | 3 | 3 |
| Omitted | LEVEL 2 DIPLOMA IN WOMENS HAIRDRESSING (TSA) | GCSE English | X | FL - fail |
| Omitted | CG 7202 LEVEL 1 DIPLOMA IN ELECTRICAL INSTALLATIONS | GCSE English | X | 3 |
| Omitted | BTEC LEVEL 2 BTEC FIRST CERTIFICATE IN BUSINESS (RG) | GCSE English | 1 | 2 |
| Omitted | LEVEL 3 SUBSIDIARY DIPLOMA IN SPORT - RUGBY ACADEMY | GCSE English | 1 | 3 |
| Omitted | AQA GCSE ENGLISH (DAY) | GCSE English | 4 | 6 |
| Omitted | BTEC EXTENDED DIPLOMA IN GAMES, ART, AND ANIMATION | GCSE English | 3 | 4 |
| Omitted | C&G 2365 DIPLOMA IN ELECTRICAL INSTALLATIONS LEVEL 2 | GCSE English | X | FL - fail |
| Omitted | LEVEL 3 OCR DIPLOMA IN ADMINISTRATION (BUSINESS PROFESSIONAL) | GCSE English | 3 | 4 |
| Omitted | CG 7202 LEVEL 1 DIPLOMA IN ELECTRICAL INSTALLATIONS | GCSE English | 3 | 2 |
| Omitted | CG LEVEL 2 NVQ DIPLOMA IN PROFESSIONAL COOKERY (TSA) | GCSE English | 2 | 3 |
| Omitted | C&G 6035 LEVEL 2 DIPLOMA IN PLUMBING STUDIES | GCSE English | U | 1 |
| Omitted | LEVEL 1 CG CERTIFICATE FOR IT USERS | GCSE English | 2 | 3 |
| Omitted | BTEC EXTENDED DIPLOMA IN 3D DESIGN | GCSE English | 1 | 2 |

| | | | | |
|---|---|---|---|---|
| Omitted | CG 7202 LEVEL 1 DIPLOMA IN ELECTRICAL INSTALLATIONS | GCSE English | 5 | 4 |
| Omitted | C&G 7202 LEVEL 1 DIPLOMA IN PLUMBING STUDIES (TSA) | GCSE English | X | FL - fail |
| Omitted | LEVEL 2 DIPLOMA IN BEAUTY THERAPY (FULL TIME) TSA | GCSE English | 5 | 4 |
| Omitted | BTEC EXTENDED DIPLOMA IN FASHION DESIGN | GCSE English | 2 | 2 |
| Omitted | AQA GCSE ENGLISH | GCSE English | 6 | 3 |
| Omitted | BTEC SUBSIDIARY DIPLOMA IN MUSIC | GCSE English | 5 | 4 |
| Omitted | BTEC EXTENDED DIPLOMA IN GAMES AND APP DEVELOPMENT | GCSE English | 2 | 3 |
| Omitted | AQA GCSE ENGLISH (DAY) | GCSE English | 9 | 8 |
| Omitted | BTEC LEVEL 2 CERTIFICATE IN PRINCIPLES OF BUSINESS ADMINISTRATION | GCSE English | 4 | 3 |
| Omitted | LEVEL 3 DIPLOMA IN EARLY YEARS EDUCATION AND CARE | GCSE English | 4 | E |
| Omitted | LEVEL 3 DIPLOMA IN EARLY YEARS EDUCATION AND CARE | GCSE English | 4 | 2 |
| Omitted | AQA GCSE ENGLISH (DAY) | GCSE English | 2 | 2 |
| Omitted | BTEC LEVEL 3 SUBSIDIARY DIPLOMA IN IT - SOFTWARE DEVELOPMENT | GCSE English | 2 | 3 |
| Omitted | OCR LEVEL 2 DIPLOMA IN IT | GCSE English | 2 | 3 |
| Omitted | BTEC EXTENDED DIPLOMA IN CONSTRUCTION | GCSE English | 3 | 5 |
| Omitted | BTEC EXTENDED DIPLOMA IN 3D DESIGN | GCSE English | 2 | 4 |
| Omitted | CORE MATHS | GCSE English | 4 | 5 |
| Omitted | CACHE LEVEL 2 EXTENDED DIPLOMA IN HEALTH AND SOCIAL CARE | GCSE English | 2 | 4 |
| Omitted | LEVEL 2 DIPLOMA IN HAIR AND MEDIA MAKE-UP (TSA) | GCSE English | 2 | 3 |
| Omitted | CITY & GUILDS DIPLOMA FOR LEGAL SECRETARIES (LEVEL 2) (LEGAL 1) | GCSE English | 2 | 3 |
| Omitted | CACHE LEVEL 1 DIPLOMA IN CARING FOR CHILDREN | GCSE English | X | 2 |
| Omitted | LEVEL 2 DIPLOMA IN BEAUTY THERAPY (FULL TIME) TSA | GCSE English | 3 | FL - fail |
| Omitted | AQA GCSE ENGLISH | GCSE English | 5 | 5 |
| Omitted | LEVEL 2 DIPLOMA IN PREPARING FOR FURTHER STUDY IN HEALTH, SOCIAL CARE AND SOCIAL WORK | GCSE English | 6 | 5 |
| Omitted | EDEXCEL GCSE MATHS (DAY) | GCSE English | 5 | 4 |
| Omitted | CITY & GUILDS DIPLOMA FOR LEGAL SECRETARIES (LEVEL 2) (LEGAL 1) | GCSE English | X | FL - fail |
| Omitted | CACHE LEVEL 2 EXTENDED DIPLOMA IN HEALTH AND SOCIAL CARE | GCSE English | 4 | 3 |
| Omitted | EDEXCEL GCSE MATHS (DAY) | GCSE English | X | 2 |
| Omitted | LEVEL 3 DIPLOMA IN SPECIALIST SUPPORT FOR TEACHING AND LEARNING IN SCHOOLS | GCSE English | X | 2 |
| Omitted | BTEC EXTENDED DIPLOMA IN CONSTRUCTION | GCSE English | 3 | 5 |
| Omitted | AQA GCSE ENGLISH (DAY) | GCSE English | X | FL - fail |
| Omitted | BTEC SUBSIDIARY DIPLOMA IN CONSTRUCTION | GCSE English | 2 | 3 |
| Omitted | BTEC EXTENDED DIPLOMA IN FILM, TV, AND SPECIAL EFFECTS | GCSE English | 4 | 2 |
| Omitted | BTEC SUBSIDIARY DIPLOMA IN CONSTRUCTION | GCSE English | 1 | 2 |

| | | | | |
|---|---|---|---|---|
| Omitted | BTEC LEVEL 2 CERTIFICATE IN PRINCIPLES OF BUSINESS ADMINISTRATION | GCSE English | 3 | 3 |
| Omitted | OCR LEVEL 2 TECHNICAL DIPLOMA IN SPORT | GCSE English | 2 | 3 |
| Omitted | BTEC LEVEL 3 SUBSIDIARY DIPLOMA IN IT - SOFTWARE DEVELOPMENT | GCSE English | 3 | 4 |
| Omitted | BTEC EXTENDED DIPLOMA IN FILM, TV, AND SPECIAL EFFECTS | GCSE English | U | 4 |
| Omitted | BTEC EXTENDED DIPLOMA IN CONSTRUCTION | GCSE English | 2 | 4 |
| Omitted | BTEC LEVEL 2 BTEC FIRST CERTIFICATE IN BUSINESS (RG) | GCSE English | 1 | 3 |
| Omitted | LEVEL 2 DIPLOMA IN BEAUTY THERAPY (FULL TIME) | GCSE English | X | 3 |
| Omitted | CACHE LEVEL 2 EXTENDED DIPLOMA IN HEALTH AND SOCIAL CARE | GCSE English | 2 | 3 |
| Omitted | AQA GCSE ENGLISH | GCSE English | 4 | 5 |
| Omitted | LEVEL 3 SUBSIDIARY DIPLOMA IN SPORT - RUGBY ACADEMY | GCSE English | 1 | 3 |
| Omitted | CG 7202 LEVEL 1 DIPLOMA IN ELECTRICAL INSTALLATIONS | GCSE English | 4 | 3 |
| Omitted | NCFE LEVEL 2 DIPLOMA FOR ENTRY TO THE UNIFORMED SERVICES | GCSE English | 3 | 3 |
| Omitted | AQA GCSE ENGLISH (DAY) | GCSE English | 5 | 4 |
| Omitted | EDEXCEL BTEC LEVEL 3 EXTENDED DIPLOMA IN APPLIED SCIENCE (FORENSIC) | GCSE English | 2 | 4 |
| Omitted | BTEC LEVEL 2 CERTIFICATE IN PRINCIPLES OF BUSINESS ADMINISTRATION | GCSE English | X | FL - fail |
| Omitted | BTEC LEVEL 3 SUBSIDIARY DIPLOMA IN IT - SOFTWARE DEVELOPMENT | GCSE English | 3 | 3 |
| Omitted | BTEC LEVEL 3 SUBSIDIARY DIPLOMA IN IT - SOFTWARE DEVELOPMENT | GCSE English | 3 | 3 |
| Omitted | C&G 7202 LEVEL 1 DIPLOMA IN PLUMBING STUDIES (TSA) | GCSE English | 5 | 4 |
| Omitted | LEVEL 2 CERTIFICATE IN AN INTRODUCTION TO EARLY YEARS EDUCATION AND CARE | GCSE English | X | 2 |
| Omitted | AQA GCSE ENGLISH | GCSE English | 8 | 5 |
| Omitted | C&G 7202 LEVEL 1 DIPLOMA IN PLUMBING STUDIES | GCSE English | 6 | 3 |
| Omitted | LEVEL 3 SUBSIDIARY DIPLOMA IN SPORT - RUGBY ACADEMY | GCSE English | 7 | 5 |
| Omitted | AQA GCSE ENGLISH (DAY) | GCSE English | X | FL - fail |
| Omitted | ACCESS TO H.E DIPLOMA (TEACHING) | GCSE English | 6 | 3 |
| Omitted | BTEC EXTENDED DIPLOMA IN VEHICLE TECHNOLOGY (MOTORSPORTS) (TECH BACC) | GCSE English | 5 | 2 |
| Omitted | TRAINEESHIP - RM9 5NU | GCSE English | 2 | 3 |
| Omitted | AQA GCSE ENGLISH | GCSE English | X | C |
| Omitted | LEVEL 3 DIPLOMA IN CHILDCARE AND EDUCATION | GCSE English | 5 | 3 |
| Omitted | CG LEVEL 2 NVQ DIPLOMA IN PROFESSIONAL COOKERY (TSA) | GCSE English | 2 | 3 |
| Omitted | BTEC LEVEL 2 BTEC FIRST CERTIFICATE IN BUSINESS (RG) | GCSE English | 2 | 3 |
| Omitted | AQA GCSE ENGLISH (DAY) | GCSE English | X | FL - fail |
| Omitted | DIPLOMA IN CARPENTRY & JOINERY LEVEL 2 | GCSE English | 2 | 3 |

| | | | | |
|---|---|---|---|---|
| Omitted | CG 7202 LEVEL 1 DIPLOMA IN ELECTRICAL INSTALLATIONS | GCSE English | X | 3 |
| Omitted | C&G 7202 LEVEL 1 DIPLOMA IN PLUMBING STUDIES | GCSE English | 5 | 3 |
| Omitted | AQA GCSE ENGLISH | GCSE English | X | 4 |
| Omitted | EDEXCEL GCSE MATHS (DAY) | GCSE English | 2 | 2 |
| Omitted | C&G 6035 LEVEL 2 DIPLOMA IN PLUMBING STUDIES | GCSE English | 1 | 3 |
| Omitted | C&G 6035 LEVEL 2 DIPLOMA IN PLUMBING STUDIES | GCSE English | 5 | 3 |
| Omitted | LEVEL 2 DIPLOMA IN HAIR AND MEDIA MAKE-UP (TSA) | GCSE English | 4 | 3 |
| Omitted | BTEC SUBSIDIARY DIPLOMA IN CONSTRUCTION | GCSE English | 4 | 4 |
| Omitted | BTEC SUBSIDIARY DIPLOMA IN CONSTRUCTION | GCSE English | 3 | 5 |
| Omitted | AQA GCSE ENGLISH | GCSE English | 6 | 4 |
| Omitted | CG 7202 LEVEL 1 DIPLOMA IN ELECTRICAL INSTALLATIONS | GCSE English | X | 1 |
| Omitted | LEVEL 3 OCR DIPLOMA IN ADMINISTRATION (BUSINESS PROFESSIONAL) | GCSE English | X | FL - fail |
| Omitted | BTEC LEVEL 2 CERTIFICATE IN PRINCIPLES OF BUSINESS ADMINISTRATION | GCSE English | 4 | 3 |
| Omitted | CG 7202 LEVEL 1 DIPLOMA IN ELECTRICAL INSTALLATIONS | GCSE English | 5 | 2 |
| Omitted | CG 7202 LEVEL 1 DIPLOMA IN ELECTRICAL INSTALLATIONS | GCSE English | 1 | 2 |
| Omitted | BTEC LEVEL 2 BTEC FIRST CERTIFICATE IN BUSINESS (RG) | GCSE English | U | 2 |
| Omitted | AQA GCSE ENGLISH | GCSE English | 2 | 4 |
| Omitted | BTEC EXTENDED DIPLOMA IN CONSTRUCTION | GCSE English | 3 | 5 |
| Omitted | LEVEL 2 CERTIFICATE IN AN INTRODUCTION TO EARLY YEARS EDUCATION AND CARE | GCSE English | 6 | 6 |
| Omitted | BTEC LEVEL 2 BTEC FIRST CERTIFICATE IN BUSINESS (RG) | GCSE English | 3 | 3 |
| Omitted | EDEXCEL GCSE MATHS (DAY) | GCSE English | 5 | 3 |
| Omitted | CORE MATHS | GCSE English | 5 | 4 |
| Omitted | LEVEL 3 OCR DIPLOMA IN ADMINISTRATION (BUSINESS PROFESSIONAL) | GCSE English | X | FL - fail |
| Omitted | SUBSIDIARY DIPLOMA IN SPORT (DEVELOPMENT, COACHING AND FITNESS) | GCSE English | X | 2 |
| Omitted | C&G 6035 LEVEL 2 DIPLOMA IN PLUMBING STUDIES | GCSE English | X | 2 |
| Omitted | AQA GCSE ENGLISH (DAY) | GCSE English | 2 | 3 |
| Omitted | LEVEL 2 DIPLOMA IN BEAUTY THERAPY (FULL TIME) TSA | GCSE English | 3 | 3 |
| Omitted | BTEC SUBSIDIARY DIPLOMA IN MUSIC | GCSE English | 3 | 3 |
| Omitted | AQA GCSE ENGLISH | GCSE English | X | FL - fail |

## 8.4 Creative writing task

**Section B: Writing**

You are advised to spend about 45 minutes on this section.
Write in full sentences.
You are reminded of the need to plan your answer.
You should leave enough time to check your work at the end.

0 5

Either:

Describe a winter's day as suggested by this picture:

Or:

Write a story about a time when you saw something for the first time.

(24 marks for content and organisation

16 marks for technical accuracy)

(40 marks)

## 8.5 Data collection methods summary

**Method A** ***(aligned to the main research question, and sub-question 2)***

***Method A...1st data capture workshop***

- Seven GCSE English teachers to undertaken comparative judgements across a sample of sixteen creative writing scripts. Each will be asked to perform 110 judgements.

***Method A...2nd data capture workshop***

- Six GCSE English teachers to contribute a sample of three creative writing pieces from their learners, contributing to a total sample of eighteen pieces. These eighteen pieces will be comparatively judged by the team together in a workshop session.
- Each teacher will complete as many comparative judgements they can in one hour from the sample of eighteen texts.

**Method B** ***(aligned to the main research question, and sub-question 1 & 2)***

Semi-structured interviews with six GCSE English teachers.

***Questions:***

1. How many years have you taught GCSE English in a Further Education setting?
2. What formal training, if any, have you participated in in teaching and assessing GCSE English? How effective was this?

3. What informal training, if any, have you participated in in teaching and assessing GCSE English? How effective was this?

---

4. What is your experience of assessing creative writing through adaptive comparative judgement?
5. Did this approach to assessment change the way you viewed each script?
6. What have you gained through assessing with comparative judgement?

---

** introduce two creative writing texts that the teacher will comparatively judge. Ask teacher to narrate the thinking they're undertaking in judging these two scripts. These scripts will be similarly ranked pieces from a previous sample**

7. Which script is more proficient as a piece of creative writing?
8. Describe what is helping you make this judgement? What are you drawing on?

---

9. Do you have any other comments you'd like to make with reference to adaptive comparative judgement?

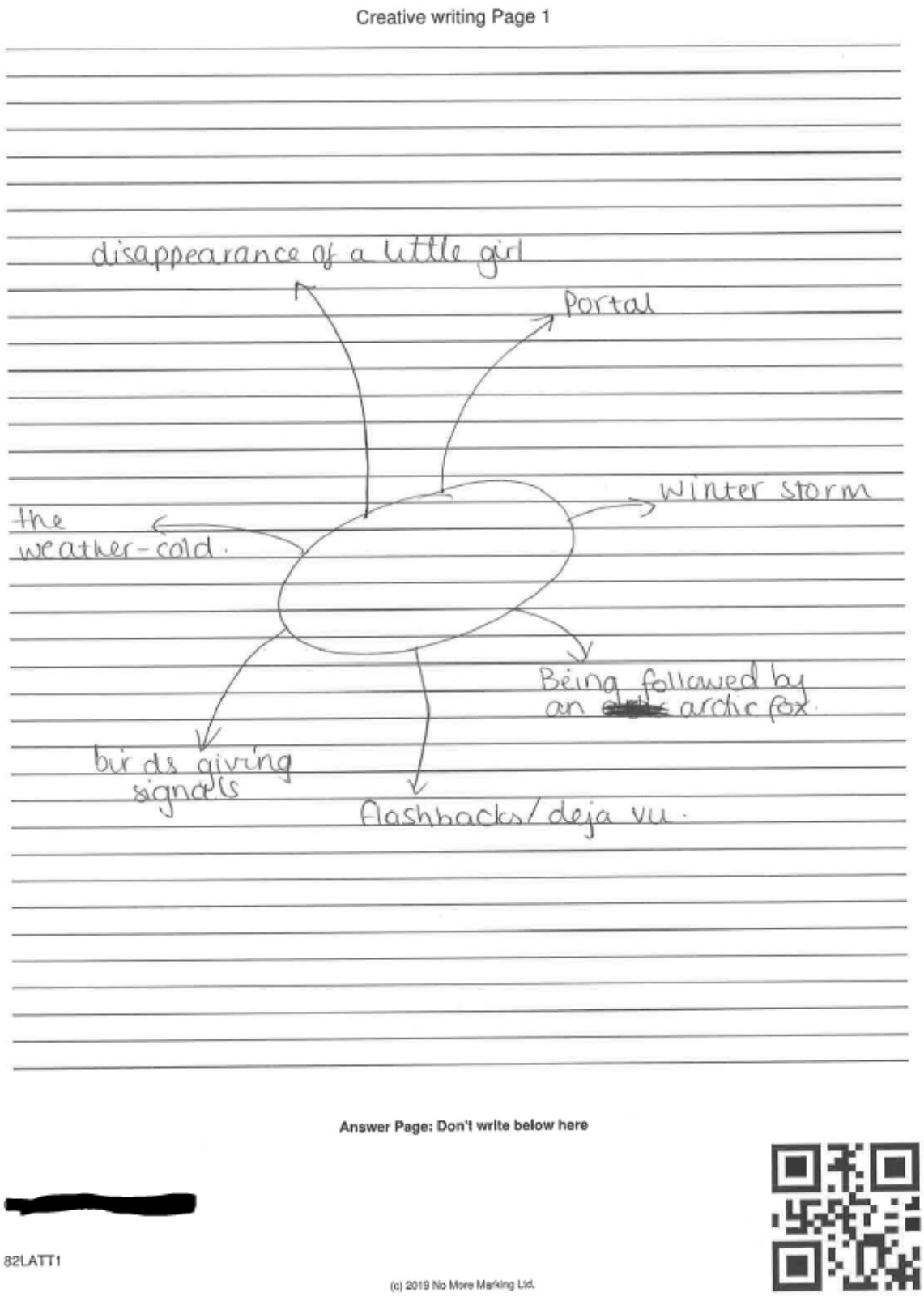
**Method C** ***(aligned to the main research question, and sub-question 3)***
Adaptive Comparative Judgement workshop with GCSE English learners.

- 10 learners to participate in comparative judgement workshop outside of normal class time.
- The sample of 18 texts gathered for method A's 2nd data capture workshop will be used.
- Learners will complete as many comparative judgements they can in one hour from the sample of eighteen texts.
- Subsequent interviews with learners (to be conducted in 2s and 3s) to follow as part of the session.

***Questions:***

- How did you decide what the better piece of writing was?
- What helped you decide?

## 8.6 Sample student creative writing script using NoMoreMarking software



I stumbled through the abandoned jungle as the trees began to close in, the icecles began to fall from the branches of the broken trees, the birds began to scatter, the ground began to shake and the little girl got closer. From where? She came? I turn to her gazing into her golden eyes as she tilts her head slightly to the left and her arm rises, pointing towards the west. I turn to the west as I follow. leading me to an cave beside a lake.

Well at least I thought it was a lake. She walked slowly like a snail towards it. I took a deep breath and swalled what felt like a but of food blocking my windpipe.

Suddenly, the trees began to close in more and the ~~girl came~~ girl became slightly more concerning, I couldn't move it was almost like I was frozen. She gave me a signal to come to her but I just couldn't. The ~~leaves to~~ icecles got bigger, the place became smaller and the girl became stranger. She just went straight to the lake and sank in almost like a portal. I gradually began to unfreeze and walked slowly towards the lake. I looked over the lake and saw spirals of blue, ~~and~~ purple and green. I thought I was going crazy at first but it was a portal. I saw her clawed, dry, and croaky hands just poking out and didn't know what to do.

Out of no where came a mysterious creature. Almost like a fox but from the Arctic. It came running out of no where howling. It was like it was telling me to run and jump before it's too late. But too late for what? I took a step back and breathed as if it was my last breath. Then I jumped.

**Answer Page: Don't write below here**

[redacted]

82LATT2

## 8.7 Audio recordings from teacher interviews

**(accessible through accompanying digital audio files)**

| Teacher | Date | Audio clip tags | Questions |
|---|---|---|---|
| 8 | 16.05.19 | Teacher 8 - ACJ 1.m4a | 1, 2, 3 |
| | | Teacher 8 - ACJ 2.m4a | 4, 5, 6 |
| | | Teacher 8 - ACJ 3.m4a | 7, 8, 9 |
| 9 | 16.05.19 | Teacher 9 - ACJ 1.m4a | 1, 2, 3 |
| | | Teacher 9 - ACJ 2.m4a | 4, 5, 6 |
| | | Teacher 9 - ACJ 3.m4a | 7, 8, 9 |
| 10 | 28.05.19 | Teacher 10 - ACJ 1.m4a | 1, 2, 3 |
| | | Teacher 10 - ACJ 2.m4a | 4, 5, 6 |
| | | Teacher 10 - ACJ 3.m4a | 7, 8, 9 |
| 11 | 29.05.19 | Teacher 11 - ACJ 1.m4a | 1, 2, 3 |
| | | Teacher 11 - ACJ 2.m4a | 4, 5, 6 |
| | | Teacher 11 - ACJ 3.m4a | 7, 8, 9 |
| 12 | 15.05.19 | Teacher 12 - ACJ 1.m4a | 1, 2, 3 |
| | | Teacher 12 - ACJ 2.m4a | 4, 5, 6 |
| | | Teacher 12 - ACJ 3.m4a | 7, 8, 9 |

## 8.8 Audio recording from student interviews

**(accessible through accompanying digital audio files)**

| Audio clip tag | Date | No. of students interviewed |
|---|---|---|
| Learnerinterview1.m4a | 04/02/2019 | 1 |
| Learnerinterview2.m4a | 04/02/2019 | 1 |
| Learnerinterview3.m4a | 16/01/2019 | 2 |
| Learnerinterview4.m4a | 22/11/2019 | 5 |
| Learnerinterview5.m4a | 22/11/2019 | 4 |

## 8.9 Student interview transcription excerpt

**Students B, C, G & H:**
**How did you decide which the better piece of writing was?**
Student G: There was one I read, I just read the first paragraph and I was like "wow", you know.
Interviewer: What was the "wow" moment?
Student G: It was just the descriptive...the language devices employed, and stuff like that
Student B: and I think the way it starts too, the attention
Student G: yeah, your attention

---

Student C: you know what puts you off? The messy writing. You don't even want to read, you just block it
Student G: But we we're not judging the style of the writing, just the writing
Student C: No, but the messy writing puts you off

---

Student H: In descriptive writing it's how they're using the words, metaphors, it's nice to read it (sic)
Student B: I think the last paragraph and the first paragraph is part of it...you think which one's good or not, that's the decider

---

Student G: when I was reading it you picture yourself there. There was one talking about the clouds making animals and I'm just there trying to visualise it to see what they're writing about
Interviewer: and was that one a good text?
Student G: I asked "why didn't I come up with that myself?"

# 8.10 Coding of student interview excerpt