Key stage 2 writing moderation

Observations on the consistency of moderator judgements

Ofqual/18/6358

Authors

This report was written by Benjamin M. P. Cuff, Emma Howard, Rebecca Mead, and Paul E. Newton, from Ofqual's Strategy, Risk and Research directorate.

Acknowledgements

We would like to thank all those who helped us during the recruitment phase, and when arranging school visits, particularly those at STA and local authority moderation managers.

We would also like to thank all the teachers and moderators who agreed to take part in this research, for sharing their views and experiences with us.

Finally, we would like to thank those who reviewed and commented on earlier versions of this manuscript, with particular thanks to Ofqual's Research Advisory Group.

Contents

Summary

At the end of key stage 2 (KS2), writing is assessed by teachers, unlike reading and mathematics, which are assessed primarily via written tests. Teacher judgements are moderated by Local Authorities (LAs), overseen by the Standards and Testing Agency (STA). In 2016, interim teacher assessment frameworks (ITAFs) were introduced, based on the new national curriculum that had been launched in 2014.

Some stakeholders began to raise concerns about the consistency of moderation of KS2 writing in 2016 and about moderator standardisation arrangements. In response to these concerns, we decided to look in more detail at moderation in 2017, in particular at the consistency of moderation judgements. We combined observations in one school in each of 12 LAs with interviews of teachers, moderators and moderation managers from around 11% (17) of the 152 LAs in England.

This was specific and focused research, observing a small proportion of moderation to provide detailed insights into aspects of the validity of assessment arrangements. Our main purpose was to identify potential risks to the consistency of moderation judgements and feedback relevant information to help STA mitigate any such risks in future years. This type of small-scale observational research can be useful in helping policy-makers understand more about what may be happening so that informed choices can be made. Our observations do not provide a definitive judgement on the quality of moderation and do not provide a broad representation of national practice.

While our research did not compare 2017 with earlier years, many participants in the study commented that they thought the ITAF was better understood in 2017 than it had been in 2016. This is supported by data provided to us by STA, which also suggests an improvement in consistency of KS2 writing assessment outcomes in 2017 compared to 2016, based on an analysis of the correlation between writing teacher assessment and reading test outcomes.

Nevertheless, we identified variations in approaches taken to moderation in 2017, including different logistical arrangements, practices and understandings of ITAF-referenced moderation. On this basis, we concluded that it was likely that moderators' judgements were more inconsistent during 2017 than they could have been, and that some variations could have operated between LAs, but that it should be possible to reduce inconsistency in future years.

We therefore recommended that STA take steps to reduce risks of inconsistency for future years; informed by the analysis within this report, as well as by its own evidence gathering. We also recommended that STA should revisit the design of the standardisation test, in light of concerns expressed about its authenticity. More broadly, our observations, including that some teachers, moderators and moderation managers had not interpreted the ITAF standards as intended, suggested that it

would be appropriate to keep the approach to the assessment of writing under review.

We begin this report by setting out the key changes that STA are putting in place for 2018, relevant to the issues raised in this report. We then describe assessment and moderation arrangements for KS2 writing in 2017 and explain the approach and rationale for the project. The main body of the report describes our observations and concludes by discussing potential risks to consistency raised by those observations.

Changes related to issues raised in this research

The primary purpose of this research was to provide feedback to STA, who oversee local authority moderation, to allow them to consider the findings and make any improvements in the approach for 2018 and future years if necessary.

Based on our research, we recommended that a number of areas should be strengthened in the short term, in particular the provision of training, guidance and support to allow for greater consistency of interpretation of assessment criteria and improvements to sampling methodology to reduce the predictability of moderation. In the longer term, we recommended that the approach to the assessment of writing should be kept under review. Our findings are discussed in the main body of this report.

While we were carrying out our research, STA were already planning to make a number of changes on the basis of stakeholder feedback, consultation and reflection on experiences in 2016 and 2017.[1] The most significant initiatives that had already been planned were:

1. The development of new teacher assessment frameworks for writing. Key changes aim to provide greater clarity, for example, the 'leading line' preceding individual ITAF statements at each standard has now been clearly included as a statement in its own right and there is now a different balance between technical elements and holistic elements, such as composition.
2. Running small scale-pilots of peer-to-peer moderation of teacher assessment and comparative judgement of writing.

These initiatives have the potential to improve the validity and reliability of teacher assessments of writing for 2018 and beyond. However, changes can also introduce new uncertainties and make it more difficult to maintain consistent standards over time.[2] Clear communication, training and guidance to teachers and moderators in relation to changes will be critical to supporting effective implementation of new teacher assessment frameworks.

In addition to initiatives set out above, STA had also already intended to make a number of changes to processes, training and communications for 2018. Our research informed these intentions, which in summary were:

1. Making improvements to moderator training and guidance to:

[1] New writing frameworks were published in September 2017. The key stage 1 framework can be found here and the key stage 2 framework here.
[2] Revisions to 'mastery' or 'secure-fit' frameworks such as these inevitably change the overall standard of assessment and impact on the performance-profile (type) of pupils reaching the expected standard.

   a. support greater consistency of process nationally by emphasising that moderators should, for example, engage with teachers through the moderation process, give schools the same minimum notice of a visit, make (and give schools consistent notice of) pupil sampling decisions in line with national guidelines and take a consistent approach to production of evidence after moderation
   b. support greater consistency of judgements: for example, by emphasising the need to exclude potentially construct-irrelevant factors from judgements and supporting local authorities to share good practice
2. Ensuring that teacher and moderator guidance for 2018, along with exemplification materials, are published at the beginning of the academic year to support teachers to effectively prepare for assessments.[3]
3. Improving the responsiveness and effectiveness of communications both to teachers and moderators, for example, improving helpline and email response times.
4. Improving the authenticity of the moderator standardisation test
5. Encouraging local authorities to moderate more than the 25% minimum sample of schools each year to reduce predictability of moderation
6. Considering whether and how more feedback can be given to moderators by STA's external moderators.

STA has engaged with the detail of this research and carefully considered our feedback. The agency has responded quickly, using our findings to inform and build on changes planned for 2018. STA has also committed to reviewing the impact of the changes they have made, including the impact of the new frameworks for 2018, as we do not yet know the extent to which these changes will affect the consistency and validity of teacher assessment outcomes. We will continue to monitor STA's response to this work and the changes that are put in place for 2018 and beyond.

[3] Guidance was published in October 2017 for use in May 2018. Key stage 1 guidance can be found here, key stage 2 here. Exemplification materials for key stage 1 can be found here and for key stage 2 here.

Background

Assessment and moderation arrangements

In 2016, new interim teacher assessment frameworks (ITAFs) were introduced (STA, 2015) to support assessment of the new national curriculum launched in 2014 (DfE, 2014). Whereas the previous system had taken a 'best-fit' approach, meaning that there was some variation in terms of what pupils at each level threshold could do, the new frameworks adopted a 'secure-fit' model, meaning that all pupils working at each attainment category (now known as 'standards') should have demonstrated secure attainment in each specified element of knowledge and skill.[4]

The 2017 iteration of the ITAF outlines teacher assessment arrangements for KS2 writing for the academic year 2016/17 (STA, 2016c). It sets out a number of 'pupil-can statements', which reflect core elements of knowledge and skill outlined in the national curriculum (DfE, 2014). These statements are distributed amongst each of 3 standards: 'working towards the expected standard'; 'working at the expected standard'; and 'working at greater depth within the expected standard'. Nine statements describe working towards the expected standard, a further 9 describe working at the expected standard, and a further 3 describe working at greater depth.[5]

KS2 writing is a 'high-stakes' assessment, as outcomes form part of primary school accountability measures, alongside reading and mathematics test outcomes.[6] At the end of KS2 (i.e. the end of school year 6, age 11) teachers are required to assess each pupil according to one of these standards, taking into account the range of writing that pupils have produced in KS2 (there is no particular assessment window during which evidence must be produced; however, later pieces of writing are typically expected to provide the bulk of evidence, when pupils' writing is the most developed). Pupils must be able to demonstrate each statement across a range of different pieces of writing. Almost any piece of writing can be used as evidence for the assessment, including those produced for other subjects (e.g. work produced for science or religious studies classes). The ITAF is not intended to direct teaching, but rather to provide a framework to help teachers assess pupils' writing that has been generated as part of their normal classroom teaching of the national curriculum.

[4] New teacher assessment frameworks have been released for use in summer 2018 (STA, 2017b).
[5] Those who do not meet requirements for working towards the expected standard are deemed to be working at 'pre-key stage 2 standards'. Such pupils are assessed under a separate ITAF (STA, 2016d), and their outcomes are not moderated by LAs. These standards therefore fall outside of the scope of this research.
[6] Although each year since 2016, the Department for Education has committed not to intervene in a school based solely on KS2 writing outcomes.

As mentioned previously, moderation of teachers' assessments of writing is overseen by STA, but delivered by LAs, who are required to moderate at least 25% of schools (STA, 2016a, sec. 6.1).[7] Moderation within each LA is led by a moderation manager. LAs cannot dictate to schools how they should be setting writing tasks, or that they need to provide evidence in any particular format. Once a school has been selected for moderation, they should be given at least 48 hours' notice, before being visited by one or more moderators (depending on the size of the cohort). During the visit, moderators are expected to review a range of work produced by a sample of pupils to check the judgements made by teachers, and at least 15% of the cohort should be included in this sample (or a minimum of 5 pupils in the case of a single class cohort). This should not include pupils deemed by their teachers to be working at pre-key stage 2 standards. Moderation is expected to be a collaborative process, with moderators and teachers being engaged in a professional dialogue throughout (STA, 2016a, sec. 4.3).

All moderation takes place before the final submission deadline for judgements. This means, therefore, that there are 3 possible outcomes of moderation for each pupil. Moderators can decide to accept the teacher's judgement, change the teacher's judgement where evidence is lacking, or give the teacher the opportunity to submit more evidence of a pupil's writing to the LA before the deadline. This latter option can be taken in cases where evidence for one or more statements may be deemed insufficient to secure the teacher's judgement, where a pupil is deemed to be close to achieving the next (higher) standard, and moderators and teachers agree that the pupil will be able to demonstrate the necessary statements before the submission deadline. Depending on LA policy, this additional evidence may or may not need to be reviewed by a moderator, before the final submission of judgements takes place. A final option that moderators can take is to expand the sample (ie to review work from a greater number of pupils), where moderators have particular concerns about the accuracy of the teacher's judgements.

Part of the STA's oversight of this process involves externally moderating LA judgements and processes. 'Operational external moderators' (OEMs) visit a number of LAs. At each, they observe a moderation visit in the morning and discuss organisational matters with the LA moderation manager in the afternoon. OEMs are required to write a report for STA, but cannot give advice or guidance to moderation managers during the visit. OEMs themselves are sometimes moderated by 'Senior External Moderators' (SEMs).[8]

[7] Including 25% of any academies that have chosen that local authority to be their moderation provider.
[8] Further information on the roles of OEMs and SEMs has been published by STA (2016b, 2017a).

Rationale and approach taken

During 2016, the first year in which new ITAFs were used, some stakeholders raised concerns about the moderation of writing at KS2, in particular in relation to consistency (and therefore the accuracy) of outcomes. For example, some head teachers suggested that there seemed to be variations in how different LAs were delivering moderation, and that some LAs were applying assessment criteria more severely than others (TES, 2016).

Stakeholders produced analyses that also seemed to suggest variations. For example, Education Datalab compared writing with reading assessment outcomes (Allen, 2016), showing that some LAs had quite high outcomes for writing relative to their outcomes for reading, whereas some had the opposite, whilst the remainder had similar outcomes for writing and reading. These variations were attributed to differences in the approach to moderation within different LAs, with some being "too harsh" and others being "too generous" in their application of the assessment criteria. While this analysis is not necessarily indicative of problems with writing moderation, as differences in reading and writing outcomes could be due to any number of factors, it and other analysis suggested that further scrutiny could be beneficial.[9]

For 2017, STA took steps aiming to strengthen arrangements, in particular, updating the ITAF and providing further guidance and national training (aimed at lead moderators). STA also introduced a new moderator standardisation test to provide additional quality assurance. This was an online test, in which moderators were given portfolios of work from three different pupils and asked to assess each portfolio to a standard (working towards, expected standard or greater depth). Moderators were required to assess all three portfolios correctly in order to pass the test. While the majority (90%) of lead moderators passed this test, two-thirds of moderators did not. Those who had only failed one of the three tasks were given further training before being approved to moderate (TES, 2017); those who had failed more than one task were not approved to moderate.

These factors, taken together with the fact that writing is used as part of school accountability measures, suggested that potential risks to the consistency of moderation would benefit from further consideration. A combination of observations and interviews had the potential to provide insights into the kinds of factors that may have had an impact on consistency, either between moderators or between different LAs. Although this kind of approach would not provide definitive conclusions about the extent or nature of any potential inconsistencies, it had the potential to generate hypotheses about the kinds of factors that could impact on judgemental consistency,

[9] See Tidd (2017), who explored variations in the differences between moderated schools and non-moderated schools in different LAs in 2016.

which could help address any risks in future. We considered three key areas: firstly, evidence of possible inconsistency between moderators; secondly, evidence of possible inconsistency between LAs and thirdly, potential causes of any inconsistency.

As part of this project, we spoke to representatives from around 11% (17) of the 152 LAs in England, and observed moderation visits in 8% of LAs (one school in each of 12 LAs)[10]. Because moderators attended multiple school visits and moderation managers tended to have insights into the practise of other LAs, the coverage of this sample is likely to be slightly broader. To achieve a reasonable spread, we targeted a range of LAs; according to geographical location[11], outcomes on Education Datalab's analysis (referenced above),[12] and performances on the STA's standardisation test. The majority of the LAs that were invited to take part did so (17 out of 23). For 5 of the 17 LAs for which the LA moderation manager agreed to take part, we were unable to schedule an observation visit. For each of the remaining 12 LAs, an Ofqual researcher observed one of the school moderation visits that took place in each area. Schools were chosen with the help of each LA moderation manager, mostly on the basis of scheduling, taking into account any other pressures that may have been affecting the schools.[13]

After observing usual moderation proceedings during a visit, the researcher spoke to teachers and moderators that had been part of the process. Moderation managers were interviewed either over the phone, once all the visits had been completed, or in person on the day of the school visit. For the 5 LAs where we were unable to schedule an observation visit, we spoke to the moderation managers only. In total, we spoke to 63 participants: 26 teachers, 20 moderators, and 17 LA moderation managers.

Our sample size of 63 stakeholders is appropriate in relation to the numbers typically reported and recommended in the academic literature for qualitative research (eg see Boddy, 2016; Guest, Bunce, & Johnson, 2006; Mason, 2010). These articles refer to the concept of 'saturation', which is where the sample size is generally deemed to be sufficient to identify all relevant themes when the collection of new data fails to uncover new information. This was largely achieved for the current

[10] Out of the approximately 4,000 schools that are moderated for KS2 writing each year.
[11] Of the 17 LAs that took part in the project, 2 were in north east of England, 2 were in the East Midlands, 2 were in the West Midlands, 3 were in the east of England, 3 were in the south east, 1 was in the south west, and 4 were in London. The north west of England was not included, as we did not wish to place additional burden on any schools that may have been affected by the incident in Manchester on 22 May 2017.
[12] We targeted some that had been labelled 'generous', some that had been labelled 'harsh', and some that had received neither label on the Datalab analyses.
[13] To avoid placing additional burden on teachers/moderators, we avoided any schools where STA external moderators (ie OEMs/SEMs) were also in attendance.

research in relation to many of the findings that we have reported, suggesting that the sampling supported plausible conclusions.

To maintain confidentiality, teachers, moderators, moderation managers and schools are not named within this report. No pupils were observed or interviewed as part of this work. More information on our method is at Appendix A.

Observations

We have grouped our key observations under three headings, which relate to the main factors that could impact on the consistency of moderation:

1. the understanding of moderation (training, appropriation of standards)
2. the logistics of moderation (the sample, notice of moderation)
3. the practice of moderation (decision-making, collaboration between moderators, involvement of teachers, feedback, and the use of additional evidence).

We then set out observations relating to the moderator standardisation test.

While occasional comparisons were made with previous years by participants, our focus was on arrangements for 2017.

1. The understanding of moderation

For moderation to be consistent, moderators must have a consistent understanding of the assessment criteria. If those implementing an assessment either do not understand or agree with the approach, there is a risk of divergence, as personal understandings or preferences may be substituted for the intended assessment criteria. Training and communication have important roles to play, as they can help ensure that criteria are understood and appropriated as intended.

1.1 Training

Training was delivered via two main routes. First, two lead moderators from each LA were invited to attend a training session delivered by STA (usually the moderation manager plus one other chose to attend). Subsequently, it became the responsibility of the LAs to cascade training down to moderators (and teachers) in their respective areas.

In terms of the centralised STA training, views were mixed. Many participants made positive comments, noting that it had helped to clarify some aspects of the ITAF, especially adding greater clarity to the statement concerning shifts in formality (see the quotes below). Others however, felt that the training focussed too much on specific points such as these, rather than a more general understanding of the ITAF.

One of the more common criticisms was that the training was delivered according to a script, and attendees were not allowed to ask any questions. The intention behind this was to make sure that each session delivered by STA was identical, to promote consistency – LAs could attend one of three training sessions – but several felt that this meant attendees had to apply their own interpretations. This had knock-on

effects on local training, when moderation managers were unable to answer the same questions posed to them by their teachers/moderators. Nevertheless, some participants recognised the benefits of this approach, compared to 2016.

In terms of the training provided to moderators by LAs, most moderators and teachers generally agreed that this was well delivered. However, moderation managers did note some difficulty in delivering training as effectively as it could have been. For example, many moderation managers also noted that because guidance had often been passed down to them quite late in the academic year, they had to deliver training much later than they would have liked. Changes in guidance also meant that training had to be changed with short notice in some LAs.

There also appear to have been some differences in how LA training was delivered to teachers and moderators. While some LAs largely repeated the same training that had been delivered by STA, others expanded upon this, for example by offering more explanation about how to interpret certain ITAF statements. Some training did not seem to achieve as much depth as was delivered by STA. Some LAs also delivered training for moderators on wider issues, such as on how to manage difficult situations (e.g. being challenged on their judgements).

1.2 Appropriation of standards

Several participants noted that understanding of the ITAF standards was more inconsistent for certain statements, or elements within statements. The statement requiring shifts in formality in writing[14] seemed to be one of the less well understood concepts, with some moderators not realising that such shifts must be well managed/controlled by the pupil in order to meet the statement.

Other statements caused some confusion by containing multiple elements, with the weighting of each element being unclear. Moderators in some LAs appeared to focus more on certain statements, e.g. spelling, than others. Some statements within the ITAF contained qualifying words such as ‘some’ or ‘most’, and there appeared to be some confusion about what how they should be interpreted, for example, how many times a particular element of punctuation should be correctly evidenced.

Another aspect of the ITAF which did not seem to be well understood was the extent to which more holistic elements of writing should be taken into account. In particular, there seemed to be variation in the extent to which moderators acknowledged the leading line that precedes each of the statements within each standard of the ITAF:

[14] To be assessed at working at greater depth within the standard, one requirement was to manage ‘shifts between levels of formality through selecting vocabulary precisely and by manipulating grammatical structures’ (STA, 2016c)

"The pupil can write for a range of purposes and audiences" (STA, 2016c, p. 4). Moderation managers in particular noted that some moderators passed over this statement and instead focused on checklists of evidence (see also Section 3.1.1).

While moderation managers acknowledged that communications were somewhat better in 2017 compared to 2016, they still noted difficulties in gaining access to necessary information. In particular, frustrations were raised about not being able to speak to contacts at STA over the phone, and not receiving a timely email reply (at that time STA operated a 15-day response time window). This had knock-on effects on moderation managers' abilities to answer queries from teachers/moderators and to cascade effective training. Some participants noted that inconsistent or changing messages made them hesitant to put guidance into practice in case policy changed again.

Sometimes guidance had been sent to individual moderation managers (eg in response to an email), and shared through unofficial channels, but not shared at a national level. Other information was occasionally shared on social media. However, these sources of information appear to have caused some confusion, as it became difficult for stakeholders to understand which pieces of guidance were official, and which were people's personal interpretations.

Due to difficulties in accessing guidance, moderation managers sometimes felt that it was up to them to apply their own interpretation of the ITAF, and of how moderation should be delivered. To try to maintain some consistency, many drew upon personal or local networks. For example, some held discussions with other managers within the local cluster of LAs and others sought information from STA in their role as external moderators. Some were members of professional associations, for example the AAIA[15], and through this had access to peer networks.

2. The logistics of moderation

The way that moderation is operated can also impact on the consistency of moderation judgements. Logistical factors, such as notice given and sampling can impact on preparations that schools are able to make.

[15] The Association for Achievement and Improvement through Assessment is a non-profit organisation, aiming to promote assessment that supports learning (http://www.aaia.org.uk/)

2.1 Moderation sample

STA sampling guidelines were generally met in our observations[16]. These guidelines require that, for each school, moderators should review a minimum of 15% of the cohort (5 for a single class cohort). This should include pupils working at each of the three standards (STA, 2016a, sec. 4).

In schools we observed, pupils were sampled for moderation in one of two ways. In 2 of our 12 observations, pupils were selected on the day by moderators. In most cases, however, schools provided their provisional judgements to the LA before moderation took place (usually 1 to 2 weeks before). Pupils were then chosen either by the moderator(s) or moderation manager, and this list was passed to schools 24-48 hours in advance of the visit, so that teachers could prepare the relevant materials. STA's guidelines do not require a particular approach and LAs are allowed to form 'local agreements' with schools, which can cover the pre-submission of teacher judgements. We observed that while some LAs had made it clear to schools that pre-submission was optional, in other areas, this seemed to be an expectation. Moderation managers told us that LAs appreciate being sent judgements beforehand so that schools do not have to gather materials on the day of the moderation visit, which could slow the process down.

One moderator commented that this short notice period was unlikely to lead to any widespread changes to judgements, although several teachers did describe doing extra preparation for sampled pupils (ie going back through materials to make sure that their judgements were secure). Similarly, one moderator explained that this practice meant that teachers could prepare what they wanted to say, which could mean that those teachers would be more able to defend any challenges to their judgements from moderator(s).

Moderators generally chose pupils from a list of the cohort that had been collated by teachers (whether on the day or beforehand). Some schools presented this list in rank order, allowing moderators to focus their efforts on 'borderline' cases (ie those closest to being awarded the next highest or lowest standard), which some liked to do. In other cases, pupils were simply grouped into standards, and were not ranked, potentially reducing the likelihood of changes to judgements at moderation.

2.2 Amount of notice

All schools sampled for moderation were generally notified on the same date, which meant that there was variation in the amount of notice that schools received, as schools moderated at the start of the 3-week moderation window received less notice

[16] Except one visit, where very slightly less (14%) of the cohort was moderated than the recommended 15%.

than those at the end. The amount of notice generally ranged between 2 to 4 weeks (although LAs only needed to give schools 48 hours' notice; STA, 2016a, p. 8). As such, there would have been different amounts of time available to schools to generate additional pieces of writing for moderation. On the other hand, while those who were moderated later tended to have more time to prepare for their visit, they also tended to have less time to submit any additional evidence, if required.

Some teachers felt they could confidently predict whether or not they would be moderated from the beginning of the academic year. For example, if they were moderated last year, they were unlikely to be moderated again this year; whereas, if they had not been moderated for some time, moderation was more likely. Where this is the case, it may mean that teacher judgements may be more secure and less likely to change at moderation.

3. The practice of moderation

Differences in the ways in which moderation guidelines are translated into practice can also impact on the consistency of moderation judgements.

3.1 Decision making processes

Moderators we observed typically took a consistent approach to organising their moderation time: considering one pupil's work at a time, reviewing each individual piece of writing in turn until it was felt that enough evidence had been seen to make a decision (i.e. to accept or change the teacher's judgement, or to request additional evidence to be submitted at a later date). However, we did identify different approaches to reaching decisions on the basis of evidence from pupil work, which could impact on judgemental consistency.

3.1.1 Overall approach

In Section 1.2, we considered differences in the extent to which moderators seemed to acknowledge the leading line of each standard within the ITAF ("The pupil can write for a range of purposes and audiences" – STA, 2016c, p. 4). Moderation managers recognised that such differences in understanding carried over to differences in approaches taken during moderation.

In all of the visits that we observed, moderators read through each piece of writing before they addressed the checklist of statements (in some cases moderators ticked off a few statements as they went along). Where moderators worked in pairs, it was sometimes the case that one focused on the more holistic elements, while the other focused on the individual ITAF statements.

We noticed some differences in the features of writing that moderators focused upon when first reading the work. For example, some moderators first tried to get a sense of the independence of writing (writing is expected to not be heavily modelled or

scaffolded by the teacher – STA, 2016a, p. 12). Some tried to gain an overall sense of which ITAF standard the writing belonged to, others focused on the general clarity of the writing; while some concentrated on the use of vocabulary, or how engaging the writing was. It is difficult to know exactly what each moderator was considering while reading each piece of work, but moderators' descriptions suggested varied approaches to the criteria against which writing was being judged:

> - I always read a piece of work… and I have a gut feeling for where they are based on many years' experience. Then I will go back and look at the criteria.
> - First and foremost… it has to be the cohesion and the verve. Because if the piece does not make sense, it doesn't matter how many features we can tick on that sheet, you cannot award the standard.
> - I like to read and check that the children are using different vocabulary and different styles, and if not then it's obviously been very highly modelled… And also to check for creativity, atmosphere that sort of thing; whether it's well written. And then we go through the ITAF statements.
> - The teacher reads it out and I close my eyes and then you're just listening for the audience and purpose, is it engaging, has it got me interested? … The main thing for me is have they engaged me as a reader, have they met their purpose in persuading me to want to do something?

Some aspects described by moderators as a focus, for example, creativity and reader interest (which are not explicitly included within the ITAF), may suggest a departure from the intended assessment standards, toward moderators' personal beliefs about what constitutes quality writing (see Section 3.1.2). Some moderators went a step further, and asked teachers to describe each pupil to them before reading the work, such as whether the pupil enjoyed writing or reading, or what kind of personality the pupil had. These moderators felt that this helped them to understand the pupil as a writer, which helped them to evaluate their work. However, this again would suggest a departure from the ITAF standards.

3.1.2 Amount and type of evidence reviewed

Moderators were fairly consistent in terms of the amount of evidence that they decided to review for each pupil. In general, they read around 3 to 5 pieces of writing, to evaluate whether ITAF statements had been demonstrated consistently by pupils. They decided to review more pieces when necessary, such as when evidence for particular statements was missing or was inconsistently demonstrated.

The amount of evidence required depended somewhat upon the statement being assessed. More fundamental statements, as perceived by moderators (eg paragraphing), tended to be evaluated against a higher threshold than other statements. Some moderation managers noted differences between LAs in this respect.

The type of evidence reviewed tended to depend upon which pieces of work teachers presented to the moderators. In some schools, moderators generally had a free choice over which pieces of writing to look at, which sometimes also included draft workbooks (moderators seemed to have access to dedicated 'draft books' in only 3 of our 12 visits). In some, teachers had identified certain pieces of work within books for moderators to look at (moderators had the option to look at other pieces of work within those books, but generally didn't do so). In other cases, moderators were given a limited selection of separate pieces of work (eg in portfolios/folders). Arguably, when teachers present specific pieces of work to moderators, it is likely that they will choose the strongest pieces. This would not be the case when moderators have more of a free choice. It is possible, therefore, that more judgements may be changed when moderators have a free choice compared to when they don't. Whether moderators have access to draft work might also have an impact on outcomes. For example, a teacher's judgement in one of our visits was changed due to the number of spelling errors found in draft writing, which wouldn't have been the case had moderators only seen final versions of the work.

3.1.3 Independence of writing

There is a requirement on moderators to "be satisfied that pupils' writing is independent" (STA, 2016a, p. 9), and guidelines have been published by STA about what constitutes independent writing (STA, 2016a, p. 12). Moderators did seem to take this requirement into account while reviewing pupils' work and, in all but one visit, teachers were asked at the start of the session to explain under what conditions their pupils' writing had been produced, and how feedback had been given. In the one visit where this did not happen, teachers were still occasionally asked how independently pupils had produced certain pieces of writing.

Most moderators stated that they generally felt confident in their abilities to detect heavily guided pieces of writing. For example, they felt able to recognise this when several pupils had produced very similar narratives, or used similar phrases/sentence structures. Where moderators had doubts over the independence of writing, they often tended to seek additional evidence from less guided pieces of work. Nevertheless, some moderators noted the importance of trusting teachers, as it was not always possible to identify writing that had been heavily led by a teacher.

In our discussions with teachers, most said that while their school's teaching of writing throughout KS2 was not purely driven by the ITAF (other untested elements of

the national curriculum were also taken into account), writing tasks were often set with these assessment standards in mind. In particular, several teachers noted that writing tasks became increasingly targeted towards the ITAF towards the end of the academic year (ie closer to the moderation). As moderators tend to pay most attention to more recent pieces, there is some risk that materials being given greater focus during moderation may potentially be those that have been produced in a relatively less independent manner.

3.1.4 Thoroughness

There were variations in the thoroughness of moderation. On occasion, moderators became less thorough for later pieces, due to time limitations (in several of our visits, moderators needed to return to their own schools in the afternoon to teach). This suggests variations in the degree to which pupils' work was scrutinised.

There were also differences in the approximate length of time that each moderation session took (for those that we observed). For example, one visit where 13 pupils were moderated took just 15 minutes more to complete than a visit where 5 pupils were moderated (the same number of moderators attended each). In 2 cases where 8 pupils were moderated, a visit in one LA took twice as long as in the other, although the same number of moderators had attended each school. In some cases where 12 pupils were moderated, 4 moderators (working in 2 pairs) took almost as long in one visit as a single moderator working alone in another. While these observations offer only rough comparisons (various factors could affect the length of these sessions), comments by moderation managers also suggested variations in thoroughness between LAs.

3.2 Collaboration between moderators

The general expectation set by STA is that only 1 moderator should need to attend each school visit (STA, 2016a, sec. 4.1), although more may be expected to attend for schools with large cohorts. In 10 of the 12 visits that we observed, 2 or more moderators were present, although not all schools had large cohorts. Interviews with moderation managers confirmed that it tended to be the norm for most LAs in our sample to send moderators to schools in pairs.

Of those 10 visits, moderators in 7 LAs worked together in pairs to review each pupil's work. In the remaining 3 visits, moderators each reviewed different pupils, but discussed difficult cases or points of clarification with one another where necessary. Moderators working together in these ways appreciated the support that they could offer one another (including supporting each other's decision when explaining them to teachers), and felt that this ultimately made their judgements more accurate. A possible implication of this is that moderators who attend visits alone may make less secure judgements, either as they lack a second opinion on borderline cases or are

less effective where teachers challenge their decisions. Conversely, moderators who operate alone could be those who are more experienced or have greater expertise.

Two of the moderation managers that we spoke to took a different approach, and ran what is known as a 'central moderation' or 'warehouse moderation', where all moderation is held in a central venue by the LA, rather than having moderators visit each school (we did not observe any of these sessions).[17] In this approach, a number of schools are moderated at the same time, each being assigned their own moderator, with a lead moderator overseeing the session. Those that ran moderation in this way perceived benefits of moderators having their colleagues available for support and advice, and of the schools being able to see that they are receiving the same treatment as others. Greater quality assurance is also possible here, as the lead moderator can observe each moderator's processes, which may help to encourage greater consistency in LAs that have adopted this model.

3.3 Involvement of teachers

STA guidance states that, "moderator(s) must review the presented pupil's work and hold a professional discussion with the year 6 teacher(s)" (STA, 2016a, p. 6). We observed some variation in the degree to which professional discussions were held. Teachers were present throughout the whole of the visit in some of the schools we visited; whereas, in others, moderators reviewed some of the materials alone (after a short introductory discussion with the teachers), before inviting the teachers back into the room to attend the remainder of the visit. Where teachers were present for all or part of moderation, there was also some variation in the extent to which moderators involved them in the process. In some visits, the teacher was fully involved, as moderators and teachers worked cooperatively to arrive at a decision while fully discussing each pupil. In other visits, moderators mostly completed the exercise themselves, but sought points of clarification from teachers where necessary (eg to locate examples of evidence, or to clarify certain decisions).

Moderators in these instances discussed various benefits of having the teacher in the room. For example, they appreciated the fact that teachers could help locate evidence for certain ITAF statements, where they were unable to locate such evidence themselves (although teachers are not required to prepare checklists of evidence before the visit, most did do so). Moderators also saw these discussions as a good opportunity for teachers' professional development, helping them to better understand the ITAF and why decisions had been made. Those that asked the teacher to leave the room while they moderated the first few pupils did so because

[17]Our findings on this particular topic were based solely on discussions held with the moderation managers who ran them. We did not observe any of these sessions and did not gain insights from any moderators/teachers who had attended one of these sessions.

they appreciated time to familiarise themselves with the materials, before holding discussions with teachers.

In a few of our visits, teachers were not present at all while moderators were reviewing pupils' work, but were given feedback after moderation finished. In some visits, members of the school's senior leadership team attended alongside teachers while moderation was taking place, whereas in others, they only attended the feedback session.

Variations in the degree of teacher involvement have the potential to cause variations in moderation outcomes, depending on the preferred approach taken by different LAs. On the one hand, teacher involvement may help moderators to make more accurate judgements, because teachers can help them to locate evidence for certain ITAF statements that may have been missed.[18] However, on the other hand, it is possible that having teachers or head teachers present may put additional pressure on moderators to agree with teachers' judgements, potentially meaning fewer judgements may change.

3.4 Feedback

In all moderation visits we observed, after moderation decisions were made, each visit closed with a formal summative feedback session which included the head teacher or another member of the school's senior leadership team. This was in line with STA requirements. Where teachers were present during moderation, they were often also given some informal feedback after each pupil's work had been reviewed.

STA moderation guidance states that "LA external moderator(s) must not dictate what schools' evidence should look like or how it is presented for an external moderation visit. In particular, LAs should not expect portfolios or checklists of evidence" (STA, 2016a, p. 9). During our observations, moderators did not dictate to schools how they should teach or generate written evidence for assessment. However, in several visits moderators recommended setting certain writing tasks that would help to generate evidence for specific ITAF statements. One moderator asked teachers to complete checklists of evidence in the future to make moderation visits run more smoothly and teachers were concerned that this was not in line with previous advice they had received.

STA moderation guidance also requires LAs to make schools aware of the appeals process before and at the beginning of the visit. In the majority of our observations (9

[18] Although teachers who were not present during live moderation were sometimes given an opportunity during the feedback session to present any evidence that moderators may have missed, it is perhaps more difficult to do so when put on the spot, compared to when books were reviewed together with the moderator(s) during live moderation.

out of 12) the appeals process was mentioned, although not always at beginning of the visit. While the appeals process may have been communicated to schools in some cases, e.g. prior to the visit, some teachers that we spoke to were not aware of it. It is important to note, though, that ultimately none of the schools that we attended wanted to appeal any decisions made.

3.5 Additional evidence

If moderators feel that evidence is insufficient to support a teacher's judgements, they may request to expand the sample (ie ask to review evidence for a greater number of pupils). This did not happen during any of our observations. If moderators and teachers agree that a pupil close to meeting a particular standard should be able to meet the standard before the final submission deadline, they may allow further evidence to be submitted during this period. This recognises that pupils may still be developing towards the end of the academic year. We observed some instances of this, where additional evidence was either needed for a pupil to remain at their current standard (i.e. when evidence was insufficient) or when it was felt that pupils were on the cusp of moving up into the next standard. In each case, teachers were given more time after moderation to submit this additional evidence, before the submission deadline.

Some participants raised concerns that because teachers know at this point exactly what evidence is needed, this could have an impact on the independence of writing, as writing tasks may be set with clear success criteria (contrary to STA guidelines). Others explained that not all LAs required all additional evidence to be submitted for re-moderation, allowing schools to internally moderate 'minor' additional evidence, and suggested that this risked gaming.

4. Moderator standardisation test

Our interviews also provided insights into moderators' views on the standardisation test. Those we spoke to felt that the reasons why many failed this test may have had more to do with limitations of the test, than with moderators' abilities. For example, many felt that one of the portfolios of work, which had not been correctly judged by many who had moderated 2 out of 3 portfolios correctly, was a borderline case. Others were concerned that the test environment was somewhat artificial as there was no opportunity for reviewing additional materials (which would be typical for borderline cases), or for professional dialogue with colleagues/teachers, a key feature of live moderation. Some also felt that the online nature of the test was more difficult than paper-based moderation, where pupil work could be more easily compared or seen together. Several moderation managers described their moderators as 'overthinking' judgements, for example, some said moderators had expected STA to provide three portfolios each with a different outcome. These types of issues are unlikely to have had an effect on live moderation.

Discussion and conclusions

While we only observed a small proportion of moderation, within our sample we saw a number of areas of consistency. For example, moderators generally met sampling requirements, each looking at approximately 15% of pupils work. Similarly, moderators we observed all considered one pupil's work at the time, reviewing each piece of work through in turn. Moderators were also fairly consistent in the amount of evidence they reviewed, typically 3 to 5 pieces of writing. In all visits we observed, moderators took account of STA guidance on the need for writing to be independently produced and asked teachers about this. Finally, summative moderation feedback was provided to teachers and head teachers at the end of all visits we observed.

We also observed a number of differences within our sample, with the potential to impact on the consistency of moderation judgements. Again, these do not necessarily represent a national picture; it is not possible to know the extent or impact of differences we observed. Nevertheless, differences can indicate areas of risk which, if addressed, could improve consistency of judgements in future. This section now discusses in turn the three areas we considered: (a) evidence of inconsistency between moderators, (b) evidence of inconsistency between LAs and (c) possible causes of any inconsistency.

Firstly, in relation to the general consistency of judgements between moderators, we observed a number of differences of practice with the potential to impact on consistency. This is particularly the case where moderators departed from intended practices by interpreting ITAF judgements through the lens of their own beliefs about what constituted high quality writing, for example, by filtering out certain kinds of evidence or focusing too much attention on particular statements. Other, less fundamental differences which have the potential to impact on consistency between moderators include:

- the amount of notice of moderation given to schools. Advance notice allows schools to focus their preparation of materials and it could mean judgements for those schools are more secure. This is also the case if schools are able to predict years in which they will be moderated
- the presentation of information to moderators; with some schools presenting a large body of work from which moderators could browse, while others encouraged moderators to look at particular identified pieces. If these pieces of work represented the most secure evidence available, or if ranked pupil lists are not provided, this could result in fewer moderation adjustments in such schools
- differences in approaches to involving teachers in moderation meetings; with some teachers being actively involved throughout, some being present

throughout but not always actively involved, and others only being consulted at the end. On the one hand, the involvement of teachers enables them to provide the moderator with useful insights into how best to evaluate pupil work. On the other hand, it becomes more likely that moderators' judgements may be unduly influenced by teacher or head teacher views
- variations in the level of thoroughness

Secondly, we turn to consideration of differences between LAs. This project was not designed to support definitive conclusions about systematic differences in approaches between LAs, or the impact of any such differences upon cross-LA judgemental consistency. Nonetheless, we observed some differences in LA policy that had the potential to impact on judgemental consistency between different LAs:

- the degree of collaboration between moderators. Moderators in some LAs worked largely independently, even when visiting the same school; moderators in other LAs worked closely within pairs; other LAs operated a central moderation process. Collaboration, and in particular the central moderation approach, increases the potential for effective quality assurance
- the degree of transparency. Central moderation could improve transparency for schools. Transparency was also an issue in relation to informing schools of their right to appeal results, which some teachers did not seem to be aware of
- the way that additional evidence was requested, elicited and processed. For example, in some LAs, schools were allowed to internally moderate 'minor' additional evidence; meaning that there were differences between LAs in the amount of scrutiny that teachers' judgements went through

Finally, although a wide variety of factors no doubt contributed to the operational differences that we observed in 2017, two areas are worth highlighting: training and related communications; and the appropriation of the ITAF standards. While participants generally felt that both these areas had improved on the previous year, concerns remained in some respects:

- training: there were concerns that national training focused on specific points rather than a general understanding of the ITAF, that it was delivered late in the academic year, that training and communications did not address bespoke concerns, and that local training diverged from national guidance
- appropriation of the ITAF: Our observations suggested that differences in how individual moderators practised could, to some extent, be attributed to differences in the degree to which they had appropriated the STA's ITAF standards. It seemed that some managers and moderators followed the 'letter' of the ITAF standards, whilst others followed their own understanding of its 'spirit', which is likely to have resulted in judgemental inconsistency. (This also seemed to be the case for teachers, as well as moderators.) If these different

understandings were spread across an LA, via training and quality assurance, this could have resulted in cross-LA judgemental inconsistency

Conclusions

Our main purpose in this work was to identify risks to the consistency of moderation judgements across LAs, based upon observations from 2017, to support STA in mitigating any such risks for future years. As such, our observations do not provide a definitive judgement on the quality of moderation, and do not provide a broad representation of national practice.

Our observations identified a number of differences in approaches taken to moderation, including logistical arrangements, practices and understandings of ITAF-referenced moderation. On this basis, we concluded that it was possible that moderators' judgements were more inconsistent during 2017 than they could have been, and that some such variations could have operated between LAs, but that it should be possible to reduce inconsistency in future years.

We therefore recommended that STA take steps to reduce the risks of inconsistency for future cycles; informed by the analysis within the present report, as well as by its own evidence gathering. We also recommended that STA should revisit the design of the standardisation test, in light of concerns expressed about its authenticity.[19] Difficulties in the consistent appropriation of assessment criteria suggested that it would be appropriate to review the ITAF to support greater clarity and more effective appropriation.[20] More broadly, our observations suggested that it would be appropriate to keep the approach to the assessment of writing under review.

Key changes that STA is putting in place for the moderation of key stage 2 writing assessments in 2018 are described in the opening sections of this report.

[19] Changes in the approach to moderator training and standardisation for 2018 are discussed on pages 6 and 7 of this report.
[20] Prior to and during this project, the ITAFs for writing were revised and replaced with a new Teacher Assessment Frameworks for writing both at key stages 1 and 2 for use in the current academic year (2017/18).

Appendix A: Method

One researcher attended each school visit to observe normal moderation practice. To avoid disrupting usual proceedings, the researcher attended as a silent observer during this stage (with the exception of brief introductions at the start). An observation form was completed to record details of what occurred during the visit Notes were taken on the scope of the visit (for example, the approach to sampling), the nature of the discussions that took place, how moderation decisions were made, and what feedback was given.

Once usual moderation proceedings were complete, the researcher conducted 2 interviews/focus-groups: 1 with the teacher(s) involved in the visit, and 1 with the moderator(s). Occasionally, the head teacher of the school also sat in on the former. The order and duration of these interviews/focus-groups depended upon the availability of participants, to fit in with their teaching commitments. Teachers were asked about their thoughts on how moderation had been conducted and what preparation they did in advance of moderation. Moderators were asked about how they made their judgements, their training and more general questions about the moderation process. Our intention was to investigate moderation practices as comprehensively as possible, in order to gain as many insights as possible into the potential for inconsistent moderation judgements. Semi-structured interviewing was used.

Interviews with moderation managers were held after our observation visits were completed, the majority by phone. These discussions did not focus on what had occurred during our visits, but were an opportunity for moderation managers to discuss moderation at an organisational level. Managers were asked about the level of guidance they had received from STA, national and local training, the standardisation tests, their experiences of being externally moderated by STA (if applicable), and their thoughts on the ITAF.

The mean duration of the interviews/focus groups was 22 minutes for teachers (range: 12 to 40 minutes), 35 minutes for moderators (range: 16 to 63 minutes), and 32 minutes for the moderation manager interviews (range: 25 to 40 minutes).

The lead researcher analysed transcripts using a thematic analysis approach. Themes were intended to be exhaustive in their coverage of commonly discussed topics, and analysis was essentially ‘semantic’ in nature (as opposed to interpretative), aiming to capture the explicit statements made by participants. Observations related to the potential for judgemental inconsistency were grouped under a number of themes, discussed above.

Appendix B: references

Allen, R. (2016). Consistency in Key Stage 2 writing across local authorities appears to be poor. Retrieved July 11, 2017, from https://educationdatalab.org.uk/2016/09/consistency-in-key-stage-2-writing-across-local-authorities-appears-to-be-poor/

Boddy, C. R. (2016). Sample size for qualitative research. *Qualitative Market Research: An International Journal*, *19*, 426–432. http://doi.org/10.1108/QMR-06-2016-0053

DfE. (2014). *National Curriculum in England Framework Document: December 2014*. London, UK: Department for Education. Retrieved from https://www.gov.uk/government/publications/national-curriculum-in-england-framework-for-key-stages-1-to-4

DfE. (2016). Primary school accountability. Retrieved July 10, 2017, from https://www.gov.uk/government/publications/primary-school-accountability

Guest, G., Bunce, A., & Johnson, L. (2006). How Many Interviews Are Enough? *Field Methods*, *18*, 59–82. http://doi.org/10.1177/1525822X05279903

Mason, M. (2010). Sample size and saturation in PhD studies using qualitative interviews. *Forum: Qualitative Social Research*, *11*. Retrieved from http://www.qualitative-research.net/index.php/fqs/article/view/1428/3027.%5BAccessed#g5

STA. (2015). *Interim teacher assessment frameworks at the end of key stage 2: September 2015*. London, UK: Standards & Testing Agency. Retrieved from https://www.gov.uk/government/publications/interim-frameworks-for-teacher-assessment-at-the-end-of-key-stage-2

STA. (2016a). *2017 teacher assessment external moderation: key stage 2 writing - For schools and local authorities*. London, UK: Standards & Testing Agency. Retrieved from https://www.gov.uk/government/publications/teacher-assessment-moderation-requirements-for-key-stage-2

STA. (2016b). *EYFS Profile , KS1 and KS2 Operational External Moderator guidance*. London, UK: Standards & Testing Agency.

STA. (2016c). *Interim teacher assessment frameworks at the end of key stage 2: July 2016*. London, UK: Standards & Testing Agency. Retrieved from https://www.gov.uk/government/publications/2017-interim-frameworks-for-teacher-assessment-at-the-end-of-key-stage-2

STA. (2016d). *Pre-key stage 2: pupils working below the test standard - Interim teacher assessment framework July 2016*. London, UK: Standards & Testing Agency. Retrieved from https://www.gov.uk/government/publications/2017-pre-key-stage-2-pupils-working-below-the-test-standard

STA. (2017a). *Teacher Assessment External Moderation 2016 - 2017: Senior*

external moderator guidance. London, UK: Standards & Testing Agency.

STA. (2017b). *Teacher assessment frameworks at the end of key stage 2: For use in the 2017 to 2018 academic year*. London, UK: Standards & Testing Agency. Retrieved from https://www.gov.uk/government/publications/teacher-assessment-frameworks-at-the-end-of-key-stage-2

TES. (2016). Moderation problems will make it impossible to evaluate primary schools fairly, heads say. Retrieved August 8, 2017, from https://www.tes.com/news/school-news/breaking-news/moderation-problems-will-make-it-impossible-evaluate-primary-schools

TES. (2017). Exclusive: More Sats “chaos” as two thirds of moderators fail to assess pupils’ work correctly. Retrieved July 11, 2017, from https://www.tes.com/news/school-news/breaking-news/exclusive-more-sats-chaos-two-thirds-moderators-fail-assess-pupils

Tidd, M. (2017). KS2 Writing: Moderated & Unmoderated Results. Retrieved July 11, 2017, from https://michaelt1979.wordpress.com/2017/04/03/ks2-writing-moderated-unmoderated-results/

We wish to make our publications widely accessible. Please contact us at publications@ofqual.gov.uk if you have any specific accessibility requirements.

This publication is available at www.gov.uk/ofqual.

Any enquiries regarding this publication should be sent to us at:

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346