

# Report on the national moderation of Inspection Reports

May 2026

## Table of Contents

|                                                                           |    |
|---------------------------------------------------------------------------|----|
| Introduction .....                                                        | 3  |
| Executive Summary .....                                                   | 3  |
| Structure of the day and tracking continued diocesan representation ..... | 4  |
| 1. Morning session: cross-regional report moderation.....                 | 6  |
| Sample size and characteristics.....                                      | 6  |
| Methodology .....                                                         | 8  |
| Findings.....                                                             | 9  |
| Quantitative checks .....                                                 | 9  |
| Qualitative checks .....                                                  | 12 |
| Analysis of moderator feedback on report strengths and weaknesses.....    | 16 |
| Conclusions .....                                                         | 19 |
| 2. Afternoon session: Focused moderation of a select sample .....         | 20 |
| Sample size and characteristics.....                                      | 21 |
| Activity 1: Grade prediction .....                                        | 21 |
| Methodology.....                                                          | 21 |
| Findings.....                                                             | 22 |
| Conclusions .....                                                         | 25 |
| Activity 2: Afl prediction.....                                           | 26 |
| Methodology.....                                                          | 26 |
| Findings.....                                                             | 27 |
| Conclusions .....                                                         | 31 |
| Activity 3 – evaluating the quality of the focused sample .....           | 32 |
| Methodology.....                                                          | 32 |
| Findings.....                                                             | 32 |
| Conclusions .....                                                         | 36 |
| 4. Other updates and wider discussions.....                               | 37 |
| Summary of discussions .....                                              | 37 |
| Action points from the day.....                                           | 40 |
| 5. Overall conclusion .....                                               | 42 |

# Introduction

## Executive Summary

This report presents the findings of the Catholic Schools Inspectorate's national moderation of inspection reports held on 21 May 2026 at 39 Eccleston Square, London. The event brought together inspection coordinators and representatives from dioceses across England and Wales to review the quality, consistency and accuracy of inspection reports published between Summer 2025 and Spring 2026. Through a combination of national sampling, focused moderation and professional discussion, the day provided a robust evaluation of report-writing practice across the inspectorate and identified priorities for future quality assurance and inspector development.

The morning moderation activity examined a representative sample of 50 reports, amounting to 12.5% of all published reports in the period under review. These reports were moderated using both quantitative and qualitative criteria derived from Template 3 of the quality assurance handbook. The findings show that report quality continues to improve nationally. In most technical and structural aspects of report writing, including grade aggregation, compliance statements, report structure, clarity and alignment between narrative and grade, performance was stronger than in previous years. Moderators judged reports more positively in 2026 than in 2025 across all nine qualitative measures, and in nearly every area more positively than in 2024. Taken together, these findings suggest that the inspectorate's quality assurance, moderation and training processes are having a sustained and measurable impact on the quality and consistency of published reports.

At the same time, the moderation exercise identified a small number of recurring weaknesses. The most significant of these concerned Areas for Improvement (Afls), which remained the least secure aspect of report writing. Moderators found considerable variability in the extent to which Afls were precise, evidence-based and clearly derived from the report narrative. Other recurring issues included the use of advisory rather than evaluative language, insufficient emphasis on impact in leadership sections, occasional imprecision in the use of framework terminology, and inconsistency in the treatment of prayer, liturgy, chaplaincy and related concepts. These weaknesses did not undermine the overall positive picture, but they do indicate clear priorities for future training and moderation.

The afternoon session added further depth to these findings through a focused moderation of four anonymised reports. This work tested the coherence between report narrative, grades and Areas for Improvement, and asked moderators to evaluate the reports themselves as professional pieces of writing. The results showed that narrative-grade alignment is now a clear strength of report writing across the inspectorate: moderators were generally able to predict the grades awarded with a high degree of accuracy. By contrast, moderators were much less consistent in predicting the Afls that would emerge from the same reports. This

confirms that the report-writing challenge is now less about making grades intelligible and more about ensuring that improvement priorities emerge naturally, convincingly and transparently from the narrative. The strongest reports were those in which evaluative writing, framework fidelity, use of evidence and improvement priorities worked together as a coherent whole.

The wider discussions during the day reinforced these conclusions and translated them into practical next steps for the inspectorate. Participants identified future training priorities in evaluative writing, Afls, framework terminology, leadership evaluation and the distinction between chaplaincy and prayer and liturgy. They also discussed the future pattern of national moderation and training events. Overall, the report presents a very encouraging picture of national report quality: standards are improving, moderation is becoming more sophisticated, and the remaining weaknesses are relatively focused and addressable. The findings therefore provide both assurance about the current quality of inspection reporting and a clear agenda for the next phase of inspector development and quality assurance work.

#### Structure of the day and tracking continued diocesan representation

This national moderation of reports meeting sought to bring together the inspection co-ordinators, or their representatives, from all 21 dioceses. There were 20 attendees in total. Four dioceses sent two representatives, and one attendee covered the dioceses of Westminster, Northampton, and Brentwood. One representative gave unexpected and unavoidably late apologies for absence. They would have been there on behalf of the two Welsh dioceses. Consequently, the level of attendance and diocesan coverage was down on last year.

Our aim is to return to full attendance and representation at next year's national Report Moderation meeting (20 May 2027), especially as there is the intention to combine both the QA of inspections and the report moderation into one national QA meeting in 2027. It is recommended that dioceses always have a backup ready in case of the last-minute necessity to send apologies. An experienced lead inspector can serve well at these national moderation meetings if the inspection coordinator is not available. Indeed, one diocese did send a lead inspector due to the inability of the diocesan coordinator to attend the meeting.

This national moderation of reports meeting covered Summer 2025 to Spring 2026 inclusive. The moderation day was divided into two different kinds of moderation activity, the first carried out in the morning, the second in the afternoon.

|                  | <b>Date of meeting</b> | <b>Interval covered</b>              | <b>No of attendees</b> | <b>Dioceses represented</b> |
|------------------|------------------------|--------------------------------------|------------------------|-----------------------------|
| <b>2022-2023</b> | 20 April 2023          | Autumn 22 – Spring 23<br>(2 terms)   | 26                     | 22/22                       |
| <b>2023-2024</b> | 8 February 2024        | Summer 23 – Spring 24<br>(3 terms)   | 20                     | 19/22                       |
| <b>2024-2025</b> | 5 June 2025            | Summer 24 – Spring 25<br>(3 terms)   | 23                     | 21/21                       |
| <b>2025-2026</b> | 21 May 2026            | Summer 25 – Spring 2026<br>(3 terms) | 20                     | 19/21                       |

*Table 1: Historical record of moderation days, coverage and attendance*

In the morning session, attendees moderated a set of five reports from a different inspection region. The feedback from this moderation was then used to generate a general comment on the quality of inspection reports nationally, and comments on individual samples will be used to provide feedback to the lead inspectors and their sponsoring dioceses on the quality of individual reports selected for the moderation activity.

In the afternoon, all attendees worked with the same set of four report samples and were asked to focus on specific aspects of these reports in order to test a set of hypotheses in relation to report writing and to assess the consistency of moderation findings in general.

# 1. Morning session: cross-regional report moderation

At the end of every term, each dioceses sends its completed reports to the CES and the sample of reports for moderation was taken from these termly submissions for Summer 2025 – Spring 2026. From each diocese, 10% of reports were selected at random for the purposes of moderation. In some cases, slightly more than 10% of reports were selected to ensure that at least one report for a primary and secondary school in each diocese was moderated.

## Sample size and characteristics

The sample was selected to ensure coverage of each of the inspection regions and each different school type, covering a proportion of each of the overall effectiveness grades awarded. The total number of inspection reports available for the moderation period was 400. However, to ensure coverage of all relevant categories of the school, the total number of samples moderated was 50 reports, representing 12.5% of reports published in those three terms. The breakdown of the number of reports available, and the sample taken from each diocese is given in the table below:

| Diocese              | Total available |           |             |          | Sample Size |           |             |          |
|----------------------|-----------------|-----------|-------------|----------|-------------|-----------|-------------|----------|
|                      | Primary         | Secondary | All Through | Tertiary | Primary     | Secondary | All Through | Tertiary |
| Arundel & Brighton   | 13              | 6         | 0           | 0        | 1           | 1         | 0           | 0        |
| Birmingham           | 33              | 8         | 0           | 0        | 3           | 2         | 0           | 0        |
| Brentwood            | 10              | 5         | 1           | 0        | 1           | 1         | 0           | 0        |
| Cardiff-Menevia      | 14              | 2         | 0           | 0        | 1           | 1         | 0           | 0        |
| Clifton              | 9               | 1         | 0           | 0        | 1           | 1         | 0           | 0        |
| East Anglia          | 2               | 1         | 0           | 0        | 1           | 1         | 0           | 0        |
| Hallam               | 12              | 1         | 0           | 0        | 1           | 1         | 0           | 0        |
| Hexham and Newcastle | 28              | 5         | 0           | 0        | 2           | 1         | 0           | 0        |
| Lancaster            | 3               | 0         | 0           | 0        | 1           | 0         | 0           | 0        |
| Leeds                | 17              | 4         | 1           | 0        | 1           | 1         | 0           | 0        |
| Liverpool            | 33              | 7         | 0           | 0        | 2           | 2         | 0           | 0        |
| Middlesbrough        | 7               | 2         | 0           | 0        | 1           | 1         | 0           | 0        |

|             | Total available |           |             |          | Sample Size |           |             |          |
|-------------|-----------------|-----------|-------------|----------|-------------|-----------|-------------|----------|
| Diocese     | Primary         | Secondary | All Through | Tertiary | Primary     | Secondary | All Through | Tertiary |
| Northampton | 4               | 2         | 0           | 0        | 1           | 1         | 0           | 0        |
| Nottingham  | 14              | 3         | 1           | 0        | 1           | 0         | 1           | 0        |
| Plymouth    | 6               | 1         | 0           | 0        | 1           | 1         | 0           | 0        |
| Portsmouth  | 8               | 1         | 2           | 0        | 0           | 1         | 1           | 0        |
| Salford     | 38              | 7         | 0           | 0        | 3           | 2         | 0           | 0        |
| Shrewsbury  | 19              | 4         | 0           | 0        | 1           | 1         | 0           | 0        |
| Southwark   | 17              | 6         | 0           | 0        | 1           | 1         | 0           | 0        |
| Westminster | 31              | 8         | 1           | 2        | 1           | 1         | 1           | 1        |
| Wrexham     | 0               | 0         | 0           | 0        | 0           | 0         | 0           | 0        |
|             | <b>318</b>      | <b>74</b> | <b>6</b>    | <b>2</b> | <b>25</b>   | <b>21</b> | <b>3</b>    | <b>1</b> |

Of the 50 reports selected for the sample, 25 were primary schools, 21 were secondary schools, 3 were all-through schools, and 1 was a sixth form college; 13 of the primary samples had nurseries, and 14 of the secondary samples had a sixth form. 2 of the all-through schools had a sixth form.

Of the sample, 11 were from the NE region, 12 from the NW region, 14 from the SE region, and 12 from the SW region; 18 were VA schools, 27 were academies (one of which was a joint school academy), 3 were independent schools, 1 was a sixth form college, and 1 was an offshore all-through school.

Of the sample, 20 were graded outstanding/excellent, 25 were graded good, and 5 were graded requires improvement/adequate. There were no inadequate/unsatisfactory schools in the sample.

The range of coverage for the sample is given in the table overleaf.

|             | Regions |    |    |    |    | School type |         |             |       |                    |                  |              |                 | Grading               |      |                               |                           |
|-------------|---------|----|----|----|----|-------------|---------|-------------|-------|--------------------|------------------|--------------|-----------------|-----------------------|------|-------------------------------|---------------------------|
|             | Total   | NE | NW | SE | SW | VA          | Academy | Independent | Joint | Sixth form college | Off-shore school | With Nursery | With Sixth Form | Outstanding/Excellent | Good | Requires improvement/Adequate | Inadequate/Unsatisfactory |
| Primary     | 25      | 6  | 7  | 6  | 6  | 9           | 16      | 0           | 1     | 0                  | 0                | 11           | 0               | 6                     | 15   | 4                             | 0                         |
| Secondary   | 21      | 4  | 5  | 6  | 6  | 9           | 11      | 1           | 0     | 0                  | 0                | 0            | 14              | 12                    | 8    | 1                             | 0                         |
| Tertiary    | 1       | 0  | 0  | 1  | 0  | 0           | 0       | 0           | 0     | 1                  | 0                | 0            | 0               | 1                     | 0    | 0                             | 0                         |
| All Through | 3       | 1  | 0  | 1  | 1  | 0           | 0       | 2           | 0     | 0                  | 1                | 2            | 2               | 1                     | 2    | 0                             | 0                         |
| Total       | 50      | 11 | 12 | 14 | 13 | 18          | 27      | 3           | 1     | 1                  | 1                | 13           | 16              | 20                    | 25   | 5                             | 0                         |

## Methodology

Working in cross-regional groups of two or three, each group had five reports to review that were taken from regions other than their own. They used an online modified Microsoft form version of QA handbook Template 3 (*Moderation of reports checklist*) to record their findings. The form asked the moderators first to check a set of binaries for each report as follows, marking them with a tick or a cross, as follows:

1. All sections of the report have been completed according to the rubric set out and correctly formatted as outlined in the style guide.
2. Overall effectiveness and key judgement grades have been correctly calculated, using the aggregate grade descriptor in the framework.
3. The 'Compliance Statement' section has been completed, indicating why the school has been given the compliance determination it has in each of the three compliance checks.
4. What the school does well' section has been completed, containing no more than five bullets that reflect the evidence presented in the report.
5. 'What the school does well' section covers all three key judgement areas.
6. 'What the school needs to improve' has been completed, the aspects are appropriate, are reflected in the evidence reported, and there are no more than three bullets.
7. The text for each of the three judgement areas contains no more than 600 words.
8. The text for each of the three judgement areas contain a paragraph covering each of: outcomes, provision, and leadership.
9. The text for each judgement matches the numerical grade given.

10. For all judgements that are not outstanding, the text makes clear what is preventing the grade from being higher than it is.

Then, using the same form, to give a qualitative judgement using the following scale:

- Very well
- Somewhat
- Not at all

These qualitative judgements were applied to the following statements for each report:

11. Judgements clearly and objectively apply the grade descriptors as they are found in the schedule.
12. The text throughout is evaluative, not just descriptive: it judges what the school does, how well it does it and the impact it has on the pupils. It focuses on typicality, on what pupils regularly experience and the evidence for this.
13. Judgements are justified by reference to evidence.
14. The report does not simply regurgitate the text of the grade descriptors from the schedule.
15. The report uses examples to support judgements and to reflect the uniqueness of the school being inspected; the school should be able to recognise the report's authenticity.
16. Comments relate to how the school is now, not how it is anticipated to be at some future point.
17. The report avoids advisory language and speculation.
18. The report is written in the present tense, is free from unnecessary technical jargon and is written in plain English.
19. The report is free from spelling and grammatical errors and adheres to the 'Guidance on grammar and house style' from the inspection handbook

Finally, attendees were asked to provide a summary of the strengths of each report (What Went Well) and a summary of where improvements were needed (Even Better If). Together these moderation checks were used to provide an indicator of the quality and consistency of reports nationally, and to provide feedback to the individual lead inspectors, and their sponsoring dioceses, on the quality of each moderated sample to support inspector development.

## Findings

### Quantitative checks

The first half of template 3 (checks 1-10), asks moderators simply to tick or cross. It is a binary judgement. This part of the form has been in use since 2023, although not all the checks were present from the beginning. In the QA Handbook, it is suggested that Template 3 is used by the QA reader in each diocese as a way of ensuring that each report meets the moderation criteria. Therefore, each report should have been through this process once already at diocesan level, using Template 3 as a guide.

Following the cross-regional moderation, the completed Template 3s were analysed to assess how well the reports met the criteria. The outcome of this exercise is given in the table (Table 2) below and comparisons with previous years shown in the chart (Fig 1) that follows:

| Template 3 binary checks                                                                                                                                                                            | Total | Possible | %age |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|----------|------|
| <b>1. Rubric</b><br>All sections of the report have been completed according to the rubric set out and correctly formatted as outlined in the style guide.                                          | 41    | 50       | 82%  |
| <b>2. Aggregation</b><br>Overall effectiveness and key judgement grades have been correctly calculated, using the aggregate grade descriptor in the framework.                                      | 47    | 50       | 94%  |
| <b>3. Compliance</b><br>The 'Compliance Statement' section has been completed, indicating why the school has been given the compliance determination it has in each of the three compliance checks. | 42    | 50       | 84%  |
| <b>4. Strengths - rubric</b><br>'What the school does well' section has been completed, containing no more than five bullets that reflect the evidence presented in the report.                     | 47    | 50       | 94%  |
| <b>5. Strengths - coverage</b><br>'What the school does well' section covers all three key judgement areas.                                                                                         | 40    | 50       | 80%  |
| <b>6. Afls</b><br>'What the school needs to improve' has been completed, the aspects are appropriate, are reflected in the evidence reported, and there are no more than three bullets.             | 31    | 50       | 62%  |
| <b>7. Word count</b><br>The text for each of the three judgement areas contains no more than 600 words.                                                                                             | 45    | 50       | 90%  |

| Template 3 binary checks                                                                                                                             | Total | Possible | %age |
|------------------------------------------------------------------------------------------------------------------------------------------------------|-------|----------|------|
| <b>8. Text coverage</b><br>The text for each of the three judgement areas contain a paragraph covering each of: outcomes, provision, and leadership. | 50    | 50       | 100% |
| <b>9. Consistency</b><br>The text for each judgement matches the numerical grade given.                                                              | 40    | 50       | 80%  |
| <b>10. Clarity</b><br>For all judgements that are not outstanding, the text makes clear what is preventing the grade from being higher than it is.   | 43    | 50       | 86%  |

Table 2 - Template 3, questions 1-10, total completed accurately in 2025-2026 as a proportion of reports moderated

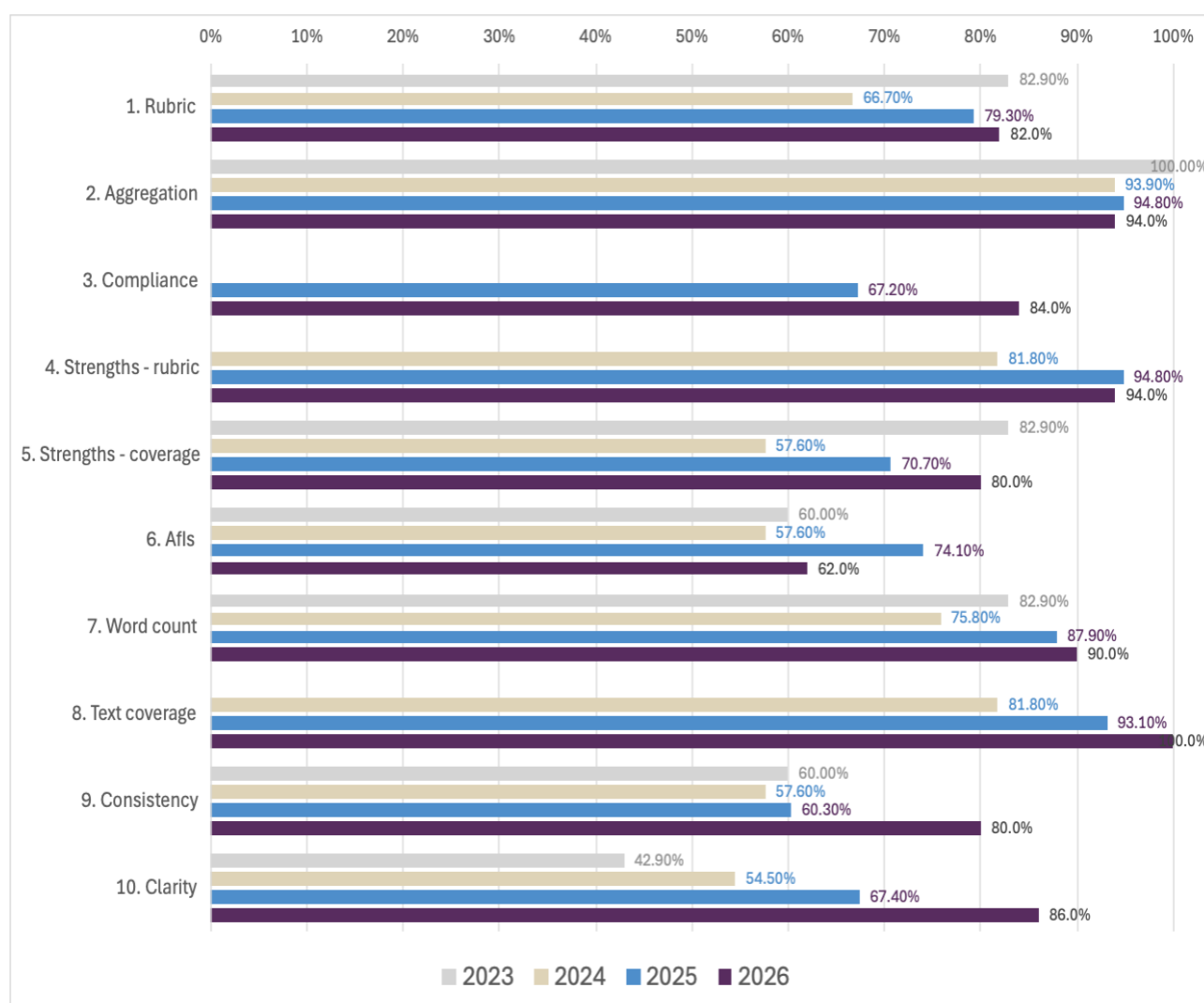


Figure 1 - Template 3, questions 1-10, %age accuracy of reports 2023-2026

As can be seen from Figure 1, the majority of reports were completed accurately and in accordance with the report-writing requirements. In almost every area, reports demonstrated improved accuracy and consistency compared with 2025, and performance was stronger across all measures than it had been in 2024. There were two areas in which scores were marginally lower than in 2025: the accuracy of grade aggregation and adherence to the report-writing rubric. In both cases, however, the decline was very small, not statistically significant, and levels of accuracy remained above 90%. These findings therefore do not indicate any meaningful deterioration in report quality.

One area, however, stood out as a notable exception to the overall pattern of improvement: the quality of Areas for Improvement (Afls). During the moderation discussion, it was suggested that the Spring 2026 reports might perform better if analysed separately, given that all inspectors had received additional training on writing effective Afls immediately prior to that inspection window. However, separate analysis of the Spring 2026 reports did not alter the overall finding. Indeed, the Spring 2026 reports performed slightly less well than the full sample, with moderators judging 47% of Afls to be accurate compared with 62% across the sample as a whole.

At face value, this could be interpreted as suggesting that the Spring 2026 training had a negative impact on inspectors' ability to write effective Afls. Such a conclusion would, however, be premature. An equally plausible explanation is that the training increased moderators' awareness of the characteristics of high-quality Afls, making them more critical and discriminating when evaluating this aspect of report writing. It is therefore difficult to determine from these findings alone whether the decline reflects a genuine reduction in quality or a change in the way quality was judged.

Nevertheless, the findings indicate that the identification and articulation of Areas for Improvement remains the least secure aspect of report writing and should continue to be regarded as a training priority. This conclusion is reinforced by the subsequent moderation activities undertaken during the afternoon session, which also identified variability in the extent to which Afls emerged clearly from the report narrative.

With this exception, the overall picture is highly positive. The analysis suggests a clear trend of improving report accuracy and consistency over the period 2024–2026, indicating that the inspectorate's quality assurance, moderation and training processes are continuing to have a positive impact on the quality of published inspection reports.

### Qualitative checks

The second half of template 3 was introduced in 2024 and asks moderators to judge a series of aspects of the reports and how well they meet the given standard, on a scale from 1 to 3, where 1 represents 'Not at all', 2 represents 'Somewhat', and 3 represents 'Very well'. The

outcome of this exercise is given in the table (Table 3) below and comparisons with 2024 and 2025 shown in the chart (Fig 2).

| Template 3 qualitative checks                                                                                                                                                                                                                                       | Raw        |          |           | Percentage |          |           |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|----------|-----------|------------|----------|-----------|
|                                                                                                                                                                                                                                                                     | Not at all | Somewhat | Very well | Not at all | Somewhat | Very well |
| <b>11. Framework</b><br>Judgements clearly and objectively apply the grade descriptors as they are found in the schedule.                                                                                                                                           | 0          | 17       | 33        | 0%         | 34%      | 66%       |
| <b>12. Evaluative</b><br>The text throughout is evaluative, not just descriptive: it judges what the school does, how well it does it and the impact it has on the pupils. It focuses on typicality, on what pupils regularly experience and the evidence for this. | 0          | 16       | 34        | 0%         | 32%      | 68%       |
| <b>13. Evidenced</b><br>Judgements are justified by reference to evidence.                                                                                                                                                                                          | 1          | 10       | 39        | 2%         | 20%      | 78%       |
| <b>14. Unique</b><br>The report does not simply regurgitate the text of the grade descriptors from the schedule.                                                                                                                                                    | 0          | 4        | 46        | 0%         | 8%       | 92%       |
| <b>15. Examples</b><br>The report uses examples to support judgements and to reflect the uniqueness of the school being inspected; the school should be able to recognise the report's authenticity.                                                                | 2          | 8        | 40        | 4%         | 16%      | 80%       |
| <b>16. Current</b><br>Comments relate to how the school is now, not how it is anticipated to be at some future point.                                                                                                                                               | 0          | 5        | 45        | 0%         | 10%      | 90%       |
| <b>17. Direct</b><br>The report avoids advisory language and speculation.                                                                                                                                                                                           | 3          | 11       | 36        | 6%         | 22%      | 72%       |

| Template 3 qualitative checks                                                                                                                                    | Raw        |          |           | Percentage |          |           |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|----------|-----------|------------|----------|-----------|
|                                                                                                                                                                  | Not at all | Somewhat | Very well | Not at all | Somewhat | Very well |
| <b>18. Plain English</b><br>The report is written in the present tense, is free from unnecessary technical jargon and is written in plain English.               | 1          | 7        | 42        | 2%         | 14%      | 84%       |
| <b>19. SPAG</b><br>The report is free from spelling and grammatical errors and adheres to the 'Guidance on grammar and house style' from the inspection handbook | 0          | 15       | 35        | 0%         | 30%      | 70%       |

Table 3 - Template 3, questions 11-19, totals for each qualitative category, with proportion of total reports moderated



Figure 2 - Template 3, questions 11-19, %age of reports at each standard 2024-2026

As can be seen from Figure 2, reports were judged more positively in 2026 than in 2025 across all nine qualitative measures. Furthermore, in eight of the nine areas evaluated, reports were also judged more favourably than those moderated in 2024. This suggests a continued improvement in the overall quality of report writing and provides further evidence that the inspectorate's moderation, quality assurance and training processes are having a positive impact.

The sole exception to this pattern concerned the extent to which reports provided direct evaluative judgements rather than advisory commentary or criticism framed as advice. Although performance in this area improved compared with 2025, it did not reach the level achieved in 2024. This finding should be interpreted with some caution. The improvement

since 2025 is encouraging, and it is possible that the comparison with 2024 reflects, at least in part, the increasing sophistication and acuity of moderators rather than a genuine decline in report-writing quality. As moderators become more experienced and familiar with the framework, their expectations of what constitutes strong evaluative writing may also become more demanding.

Nevertheless, the issue was sufficiently prominent to emerge repeatedly in moderators' qualitative comments on report quality. Across a number of reports, moderators identified a tendency for inspectors to drift into advisory language or to imply recommendations within the narrative rather than presenting clear evaluative judgements about the school's current effectiveness. This finding is consistent with observations made during previous moderation exercises and suggests that maintaining a clear distinction between evaluation and advice remains a challenge for some inspectors.

While the overall trend is one of sustained improvement, the findings indicate that the use of direct evaluative language should remain a focus of future inspector training and report moderation. Strengthening inspectors' confidence in making clear evaluative statements, without resorting to advisory language, would further enhance the consistency, clarity and professional quality of inspection reports.

Finally, moderators were asked to give a summary of the strengths (WWW) and weaknesses (EBI) of each report and to include this on the template 3 form for each report. Based on the moderator comments in the feedback forms, together with the themes emerging from both the discussions that followed the moderation exercise, several consistent findings emerge. Perhaps the most striking conclusion is that moderators were generally positive about the overall quality of the reports. While much of the discussion focused on what could be improved, there was broad agreement that report quality has continued to improve and that the vast majority of reports are fit for purpose, evidence-based and aligned with the framework. The moderation exercise therefore identified a relatively small number of recurring weaknesses against a backdrop of generally strong report-writing practice.

#### Analysis of moderator feedback on report strengths and weaknesses

##### *Overall strengths identified by moderators*

A strong theme emerging from both the written feedback and the subsequent discussion was that reports are becoming increasingly consistent in structure, tone and application of the framework. Moderators frequently commented that reports were generally clear, professionally written and capable of communicating inspection findings effectively to schools. There was widespread agreement that inspectors are increasingly confident in applying the framework and that this was reflected in the quality of evaluative writing and the consistency of judgements.

Many moderators highlighted the following strengths:

- Strong alignment between narrative and grades awarded.

- Improved consistency in applying grade descriptors.
- Generally good use of evaluative language.
- Greater clarity and coherence than in previous years.
- Effective identification of strengths and areas requiring development.
- Improved compliance with report-writing requirements.
- Reports that were generally accessible and understandable for school leaders and governors.

Particular praise was given to reports that successfully balanced celebration and challenge, recognised the uniqueness of individual schools, and used examples or quotations to bring the report to life. Moderators repeatedly noted that the strongest reports were those that demonstrated clearly how inspection evidence led to evaluative judgements and ultimately to Areas for Improvement.

There was also recognition that the introduction of new systems and processes, including use of Airtable and revised reporting templates, had contributed positively to consistency and reduced technical errors. Moderators noted that aggregation errors had virtually disappeared and that compliance statements were generally stronger than in previous years.

#### *Evaluative writing versus descriptive writing*

The most frequently identified weakness concerned the distinction between evaluation and description.

Moderators repeatedly observed that inspectors sometimes described what schools do rather than evaluating the effectiveness or impact of what they do. This issue was particularly evident within leadership sections, where reports often catalogued leadership activities without making sufficiently clear the difference those activities had made to pupils or to the Catholic life of the school.

Discussion during the moderation exercise repeatedly returned to the importance of answering the implicit 'so what?' question. Moderators argued that the strongest reports move beyond description to explain the significance and impact of evidence. Where reports failed to do this, judgements appeared less convincing and narrative sections risked becoming formulaic.

This was particularly apparent in leadership sections, where moderators frequently identified confusion between leadership activity and provision. Several reports included descriptions of what leaders had put in place but gave insufficient attention to evaluating the resulting impact.

#### *Advisory language and indirectness*

The second major theme concerned the continued use of advisory language. Moderators consistently identified the use of phrases such as:

- 'would benefit from'
- 'could be developed further'
- 'should continue to'
- 'greater consistency is required'

as a weakness in report writing.

There was widespread agreement that this type of language often obscures the underlying evaluative judgement. Rather than stating clearly what is not currently in place, reports sometimes relied upon indirect recommendations which softened the force of the judgement and made weaknesses less explicit.

Several moderators argued that advisory language creates ambiguity for schools and can make it more difficult to understand precisely why an Area for Improvement has been identified. This theme emerged repeatedly in both written comments and group discussion and was one of the most consistent findings across the moderation exercise.

#### *Areas for Improvement*

The quality of Areas for Improvement emerged as the single most significant report-writing issue.

Moderators identified several recurring concerns:

- Afls were not always clearly evidenced in the report narrative.
- Some Afls appeared more generic than specific.
- Afls occasionally lacked precision and measurability.
- Reports sometimes included Afls in all judgement areas regardless of whether this was justified by the evidence.
- In some cases moderators struggled to identify the eventual Afls from the narrative alone.

Discussion suggested that some inspectors continue to view Afls as a requirement for each judgement area rather than as a targeted response to the most significant weaknesses identified during inspection.

There was strong consensus that Afls should emerge naturally from the report narrative and should represent the most significant priorities for improvement rather than a comprehensive list of developmental suggestions.

#### *Terminology and framework understanding*

A substantial proportion of comments related to terminology.

Moderators identified continuing inconsistency in the use of terms such as:

- prayer
- liturgy
- collective worship
- celebration of the word
- assembly
- chaplaincy

The discussion highlighted that confusion over terminology often reflected deeper uncertainty about the framework itself, as well as of the new vocabulary introduced in *To Love You More Deeply*. Particular concern was expressed regarding the use of 'liturgy' as a catch-all term and the tendency to conflate chaplaincy activity with prayer and liturgy.

Moderators felt that greater consistency in terminology would improve report accuracy and reduce confusion for schools.

#### *Pupil leadership and independence*

A notable theme emerging from discussion was the interpretation of pupil leadership within prayer and liturgy.

Moderators expressed concern that some reports appeared to equate excellence with increasing levels of pupil independence. Discussion concluded that this represents an oversimplification of the framework. Participants emphasised that the quality of prayer and liturgy is more important than the degree of pupil independence and that independent pupil leadership is not, in itself, a measure of excellence. The ability of pupils to plan Celebrations of the Word independently could be an indicator of other areas the framework does highlight – such as pupil engagement and as evidence of pupil understanding, for example – but only if their independence is matched by skill in planning, ministering and evaluation prayer and liturgy. Pupil formation in planning, ministry, and evaluation of prayer and liturgy would be the key to ensuring the pupil independence in this area was a genuine marker of quality, rather than a proxy for it.

This theme appeared regularly in both written comments and verbal discussion and was identified as a future training priority.

#### *Balance and coverage*

Moderators also identified concerns regarding the balance of content across some reports.

Particular issues included:

- over-emphasis on Catholic Social Teaching;
- insufficient coverage of some framework descriptors;
- inconsistent attention to RSHE;
- uneven distribution of content between sections;
- occasional over-reliance on familiar themes at the expense of broader evaluation.

Several moderators suggested that inspectors tend to write more confidently about areas with which they are most familiar, leading to disproportionate coverage of certain themes.

## Conclusions

The moderation exercise identified a number of recurring report-writing strengths and weaknesses. The strongest reports were characterised by clear evaluative language, strong alignment between evidence and judgement, effective use of examples and quotations, and Areas for Improvement that emerged naturally from the narrative. The principal weaknesses identified were the continued use of advisory language, insufficient focus on impact in leadership sections, inconsistent use of framework terminology, and variability in the quality and precision of Areas for Improvement. Despite these issues, moderators were clear that report quality has improved significantly over recent years. The findings therefore point not to systemic weaknesses in report writing, but to a relatively small number of recurring themes that should inform future training, moderation and quality assurance activity.

## 2. Afternoon session: Focused moderation of a select sample

The afternoon session of the national moderation meeting was devoted to the focused moderation of shared set of four inspection reports. The purpose of this work was to explore the quality, coherence and effectiveness of report writing across the inspectorate by examining how moderators interpreted the same sample of published reports. Four anonymised inspection reports, representing a range of grades and contexts, were selected for moderation. Participants worked in cross-regional groups and were provided with reports from which the published Areas for Improvement (Afls) and the inspection grades had been removed. Each group completed three separate moderation activities, designed to test different aspects of report quality.

The first activity examined the relationship between the narrative of the report and the grades awarded. Moderators were asked to read each report and predict the grades that they believed had been awarded for Overall Effectiveness, Catholic Life and Mission, Religious Education and Collective Worship, together with their respective sub-judgements. Their predictions were subsequently compared with the actual grades. The purpose of this activity was to test the extent to which the evaluative language, evidence and overall tone of the report clearly communicated the judgements reached by inspectors. A high degree of agreement between moderator predictions and the published grades was taken as evidence that the report provided a clear and coherent rationale for its judgements. Conversely, if moderators struggled to predict the grades, this would suggest that the relationship between narrative and judgement may have been less secure or that the report described a genuinely borderline case.

The second activity focused on the Afls. Having read the report narrative, moderators were asked to identify the three Afls they believed would most naturally arise from the evidence and evaluation presented. These predictions were then compared with the Afls that appeared in the published report. The purpose of this exercise was not to judge whether the Afls themselves were appropriate, but to test whether they emerged convincingly from the report narrative. Reports in which moderators independently identified the same improvement priorities as the lead inspector were judged to demonstrate a strong and transparent line of reasoning from evidence to improvement priorities. Divergence between moderator predictions and the published Afls suggested either that multiple weaknesses competed for attention within the report or that the rationale for the selected Afls was less clearly signalled.

The third activity invited moderators to evaluate the reports themselves as pieces of professional writing. Groups were asked to identify what the report did well and what could be improved, drawing on the report-writing quality assurance criteria and their own professional experience. Particular attention was given to issues such as evaluative language, use of evidence, authenticity, adherence to house style, use of stakeholder voice, clarity of reasoning

and the extent to which the report reflected the uniqueness of the school. The purpose of this exercise was to identify common strengths and weaknesses in report-writing practice and to inform future inspector training and quality assurance activity.

Taken together, these three activities provided a comprehensive assessment of report quality. The first tested the coherence between narrative and grade, the second the coherence between narrative and improvement priorities, and the third the quality of the report as a professional evaluative document. Collectively, they offered valuable insight into the effectiveness of report writing across England and Wales and identified areas of strong practice as well as opportunities for further development.

### Sample size and characteristics

The four samples selected for the focused moderation were drawn from the four inspection regions and represented a range of types of school and inspector experience:

**Sample A** – A Welsh, diocesan, voluntary aided primary school from the southwest inspection region. The school had been given an overall effectiveness grade 4 (unsatisfactory in the nomenclature of the inspectorate in Wales). The lead inspector had twenty-five years' inspection experience.

**Sample B** – A diocesan, secondary academy from the northeast inspection region. The school had been given an overall effectiveness grade 2 (good). The lead inspector had six years' inspection experience.

**Sample C** – A diocesan, voluntary aided primary school from the northwest inspection region. The school had been given an overall effectiveness grade 3 (requires improvement). The lead inspector had twelve years' inspection experience.

**Sample D** – A secondary academy in religious order trusteeship. The school had been given an overall effectiveness grade 1 (outstanding). The lead inspector had five years' inspection experience.

### Activity 1: Grade prediction

#### Methodology

This moderation activity was designed to test the consistency of the selected report samples: is the narrative of the report clear enough that an educated reader or experienced moderator is able to infer the grade given without having to see it. Moderators worked in eight cross-regional groups of two or three. Four anonymised report extracts (Samples A–D) were selected and the published grades removed. Moderators were asked to read each report and to guess what grade the lead inspector had given for each sub-judgement area, and thereby also guess the grade for each key judgement area and the overall effectiveness grade. Each groups' guesses were compared with the actual grades given for each sub-judgement, each key judgement and the overall effectiveness grade. Two measures are given for each grade for each sample: the total number of groups that accurately guessed the grade and a measure of

the convergence between the groups. This latter measure was arrived at by taking an average of the groups guesses and presenting this as a percentage of the actual grade given for that key judgement area. This measured the rate of divergence from the actual grade as well as the total amount of agreement overall.

## Findings

The tables below show how well moderators were able to accurately predict the grade given for each judgement and sub-judgement for each sample based on the report narrative. Table 1 shows the actual grades awarded for each judgement and sub-judgement for each sample. Table 2 shows how many groups out of 8 were accurately able to guess the correct grade for judgement and sub-judgement for each sample. Table 3 shows a divergence measure for each guess. This measure was generated by calculating an average guessed grade and subtracting this from the actual grade. This shows how far the guesses deviated from the actual grade, as a proportion of one grade for each judgement and sub-judgement for each sample. A score of zero indicates that there was no deviation from the correct grade for this judgement or sub-judgement, ie that every group accurately predicted the grade given in that case.

|   | OE | CLM | clm1 | clm2 | clm3 | RE | re1 | re2 | re3 | CW | cw1 | cw2 | cw3 |
|---|----|-----|------|------|------|----|-----|-----|-----|----|-----|-----|-----|
| A | 4  | 3   | 3    | 3    | 3    | 4  | 3   | 4   | 3   | 3  | 3   | 3   | 3   |
| B | 2  | 1   | 1    | 1    | 1    | 2  | 2   | 2   | 1   | 2  | 2   | 2   | 2   |
| C | 3  | 3   | 3    | 3    | 3    | 3  | 3   | 3   | 3   | 3  | 3   | 2   | 3   |
| D | 1  | 1   | 1    | 1    | 1    | 1  | 1   | 2   | 1   | 1  | 1   | 1   | 1   |

Table 4: Actual grades given by lead inspectors for each judgement and each sub-judgement for each sample

|   | OE  | CLM | clm1 | clm2 | clm3 | RE  | re1 | re2 | re3 | CW  | cw1 | cw2 | cw3 |
|---|-----|-----|------|------|------|-----|-----|-----|-----|-----|-----|-----|-----|
| A | 7/8 | 7/8 | 7/8  | 3/8  | 7/8  | 7/8 | 3/8 | 5/8 | 5/8 | 7/8 | 8/8 | 8/8 | 7/8 |
| B | 7/8 | 8/8 | 8/8  | 8/8  | 8/8  | 7/8 | 7/8 | 8/8 | 7/8 | 4/8 | 4/8 | 7/8 | 3/8 |
| C | 6/8 | 3/8 | 3/8  | 3/8  | 4/8  | 6/8 | 8/8 | 8/8 | 6/8 | 8/8 | 7/8 | 1/8 | 8/8 |
| D | 6/8 | 7/8 | 5/8  | 8/8  | 7/8  | 7/8 | 6/8 | 6/8 | 7/8 | 7/8 | 4/8 | 7/8 | 7/8 |

Table 5: Number of correct guesses of grade for each judgement and sub-judgement for each sample

|   | OE    | CLM   | clm1  | clm2  | clm3  | RE    | re1   | re2   | re3   | CW    | cw1   | cw2   | cw3   |
|---|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| A | 0.125 | 0.125 | 0.125 | 0.625 | 0.125 | 0.125 | -0.63 | 0.375 | -0.38 | -0.13 | 0     | 0     | -0.13 |
| B | 0.125 | 0     | 0     | 0     | 0     | 0.125 | 0.125 | 0     | -0.13 | 0.5   | 0.5   | 0.125 | 0.625 |
| C | -0.25 | 0.625 | 0.625 | 0.625 | 0.5   | -0.25 | 0     | 0     | -0.25 | 0     | 0.125 | -0.88 | 0     |
| D | -0.25 | -0.13 | -0.38 | 0     | -0.13 | -0.13 | -0.25 | 0.25  | -0.13 | -0.13 | -0.5  | -0.13 | -0.13 |

Table 6: Average divergence from actual grade for each judgement and sub-judgement for each sample

| Rank | Sample | Average number of correct guesses | Average divergence from correct grade | Interpretation of Narrative-Grade alignment |
|------|--------|-----------------------------------|---------------------------------------|---------------------------------------------|
| 1    | B      | 6.6                               | 0.15                                  | Very strong.                                |
| 2    | D      | 6.5                               | -0.15                                 | Very strong.                                |
| 3    | A      | 6.2                               | 0.03                                  | Strong.                                     |
| 4    | C      | 5.5                               | 0.07                                  | Moderate.                                   |

*Table 7: Accuracy of moderators' grade predictions in rank order for each sample*

When looking at the four headline grades (Overall Effectiveness, Catholic Life and Mission, Religious Education and Collective Worship), and each sub-judgement grade, moderator accuracy was remarkably high across all four samples. When this is compared with the moderators' success in guess Areas for Improvement for each sample (see Activity 2 below) it is revealing. The most interesting finding is that the ranking for grade predictability differs from the ranking for Afl predictability. This suggests that a report can communicate its grades very clearly whilst still leaving readers uncertain about the eventual Areas for Improvement. An analysis of each sample's consistency between grade and narrative is given below.

*Sample A – Strong consistency between narrative and grade.*

This report produced a very high level of agreement between moderators and the lead inspector's grading:

- 7 of 8 groups correctly predicted Overall Effectiveness as a grade 4.
- 7 of 8 correctly predicted Catholic Life and Mission as a grade 3.
- 7 of 8 correctly predicted Religious Education as a grade 4.
- 7 of 8 correctly predicted Collective Worship as a grade 3.

The guesses for the sub-judgements were slightly less accurate, but on average 6 out of 8 guesses were accurate across all sub-judgements while the key judgments and overall effectiveness grade were correctly guessed by 7 out of 8 of the moderator groups. In no instance did the guess deviate from the correct grade by more than one grade, with the lowest rate of deviation between the guessed grade and the actual grade.

Moderators found guessing the grades in this sample relatively easy because the narrative contains repeated evaluative language signalling significant weaknesses, such as:

- 'limited differentiation'
- 'little evidence'
- 'expectations are generally low'

- 'self-evaluation lacks rigour'
- 'requires improvement' language appears repeatedly throughout the text.

The report makes little attempt to soften the judgement and repeatedly returns to weaknesses in teaching, leadership and evaluation. This report demonstrates a high level of consistency between narrative and grades because the evaluative nature of the report consistently signalled the lower grades, and the lowest grade are (RE) had the most strongly worded negative evaluations.

#### *Sample B – Very strong consistency between narrative and grade*

This was the strongest report for grade predictability:

- 7 of 8 correctly identified Overall Effectiveness as a grade 2.
- All groups correctly identified Catholic Life and Mission as a grade 1.
- 7 of 8 correctly identified Religious Education as a grade 2.
- The guesses for Collective Worship was split between Grade 1 and Grade 2, suggesting it was seen as borderline by moderators (and probably by the lead inspector too).

The guesses for the sub-judgements were also highly accurate in two out of three areas. In all the CLM sub-judgements all groups correctly identified the grades. In the RE sub-judgements at least 7 of 8 of groups correctly identified the grades. The CW sub-judgements were less accurately predicted, but as with sample A, never by more than one grade.

Moderators found guessing the grades in this sample straightforward because the report is explicit in its evaluative language. The Catholic Life and Mission section repeatedly uses terms such as 'outstanding', 'exceptional', 'highly valued', 'inspirational', and 'unwavering' throughout all three strands. Conversely, the RE and CW sections contain a small number of clearly identified limitations which prevent Grade 1, such as questioning not being consistently deep enough and limited parish engagement.

As a consequence, moderators could immediately see that Catholic Life and Mission was outstanding but that RE and Collective Worship were less than Outstanding. They correctly identified that RE was good and knew that CW sat at the bottom of grade 2, or the top of grade 3, as reflected in the split grades. The borderline nature of this section of the sample is reflected in the narrative evaluation for Collective Worship. This is arguably the clearest example of narrative-grade consistency in the sample set.

#### *Sample C – Moderate consistency between narrative and grade*

This sample produced the lowest level of agreement:

- 6 of 8 correctly identified Overall Effectiveness as a grade 3.
- Only 3 of 8 correctly identified CLM as a grade 3.
- 6 of 8 correctly identified RE as a grade 3.

- 8 of 8 correctly identified Collective Worship as a grade 3.

The guesses for the sub-judgements were equally mixed, with half or fewer of the groups guessing the grade correctly in 4 out of the nine sub-judgements. The deviations were again never more than one grade adrift from the actual grades, but the grades deviated more often than in the other three samples.

The moderators struggled more with this sample because the report contains a mixture of positive and negative evaluation. For example, phrases such as 'pupils enjoy lessons' and 'parish links are very strong' alongside phrases like 'weak assessment' and 'inconsistent practice'. The overall impression is more mixed than in Samples A or B. While most could identify that this was a grade 3 school throughout it appears to sit close to the Grade 2 / Grade 3 boundary. Moderators generally recognised the weaknesses but differed in how much weight they gave them. The exercise therefore suggests not that the grades were unclear, but that the report described a genuinely borderline school.

#### *Sample D - Very strong consistency between narrative and grade*

This result is perhaps the most interesting, given the contrast with the ability of moderators to accurately predict the Afls (see Activity 2 below):

- 6 of 8 groups correctly identified Overall Effectiveness as a grade 1.
- 7 of 8 correctly identified CLM as a grade 1.
- 7 of 8 correctly identified RE as a grade 1.
- 7 of 8 correctly identified CW as a grade 1.

This is a very high level of agreement. The success at guessing the sub-judgements was also very high with 8 out of the 9 sub-judgements being correctly guessed by 6 or more of the groups. This is only slightly lower than the strongest alignment by moderators seen for sample A. Again, as with all other samples the grade never deviated by more than one grade from the actual grade.

Moderators found guessing the grades in this sample straightforward because the report consistently uses the language of excellence: 'outstanding', 'exemplary', 'profound' etc. The few criticisms are clearly framed as developmental refinements rather than substantive weaknesses, referencing providing 'more opportunities' or 'greater consistency'. Moderators had little difficulty recognising that this was an outstanding report. Interestingly, this contrasts with the Afl exercise where moderator predictions diverged significantly. This was no doubt because with an outstanding school the areas for improvement are always going to be more difficult to identify and represent areas where practice will already be very strong.

#### Conclusions

The exercise suggests that, across all four reports, consistency between narrative and grade is strong. Moderators can accurately predict grades given in the majority of cases. This would

suggest that consistency between narrative and grade is a strength of report writers and, when contrasted with Activity 2 (below) grade integrity is generally stronger than Afl integrity. In other words, moderators were more successful at answering: 'What grade would I award this report?' than: 'What Areas for Improvement would I write?' This is encouraging from a grading perspective because it indicates that the evaluative language of the reports generally aligns well with the judgements awarded. The greater variability in the Afl exercise suggests evidenced in the next activity suggests that identifying and prioritising the most significant improvements remains the more challenging aspect of report writing.

## Activity 2: Afl prediction

### Methodology

This moderation activity was designed to explore the relationship between the narrative content of inspection reports and the Afls that arise from them. Moderators worked in eight cross-regional groups of two or three and undertook a report-quality exercise designed to examine the extent to which Afls emerges clearly and convincingly from the narrative of inspection reports. Four anonymised report extracts (Samples A–D) were selected and the published Afls removed. Moderators were asked to read each report and identify what they believed the Afls would have been based solely on the evidence and evaluation contained within the report text. Their predictions were then compared with the actual Afls published in the final reports, and a score given for each one of the group's predicted Afls for each sample, using the following scale:

- 0 - no correspondence of any kind between the prediction and any of the actual Afls; a miss.
- 1 - the predicted Afl was in the same broad inspection area as at least one of the actual Afls; a low-level match
- 2 - the predicted Afl was in the same inspection area and partially matched the specific improvement identified in at least one of the actual Afls; a partial match
- 3 - the predicted Afl was a close or direct match with at least one of the actual Afls

The scores for each predicted Afl were added together for each group's predictions and a score for each sample generated by providing a mean average for these group prediction totals (out of a possible perfect score of 9). In addition, a score out of three was given for each group's guesses which corresponded to how many of their predicted Afls scored a partial match with at least one of the Afls, and these in turn were averaged for each sample.

The purpose of the exercise was not to evaluate the quality of the Afls themselves, but to test the coherence between the evidence presented, the evaluative judgements made by inspectors and the improvement priorities ultimately identified. Reports in which moderators were able to predict the published Afls with a high degree of accuracy were taken to demonstrate a strong and transparent line of reasoning from evidence to improvement priorities. Conversely, where moderator predictions diverged significantly from the published

Afls, this suggested either that multiple weaknesses competed for attention within the report or that the rationale for the chosen Afls was less clearly signalled. The exercise served as a valuable quality assurance measure of report-writing effectiveness, testing whether the most significant findings of an inspection are communicated with sufficient clarity and emphasis to lead moderators independently to the same conclusions as the report author. The exercise therefore provided an indirect but robust measure of report-writing quality, particularly the extent to which Afls emerged naturally and convincingly from the inspection evidence presented.

## Findings

The scores for each sample are given in the table below and an analysis of each report's scores follows it.

| Rank | Sample | Prediction accuracy (/9) | No. of accurate predictions (/3) | Interpretation                                          |
|------|--------|--------------------------|----------------------------------|---------------------------------------------------------|
| 1    | A      | 6.13                     | 2.3                              | Strongest correlation between predicted and actual Afls |
| 2    | B      | 3.63                     | 1                                | Some clear predictability, but only in one area         |
| 3    | C      | 3.25                     | 1.1                              | Moderate to weak predictability                         |
| 4    | D      | 2.38                     | 0.5                              | Weakest correlation between predicted and actual Afls   |

Table 8: Accuracy of moderators' Afl predictions in rank order for each sample

### Sample A – Strongest correlation between predicted and actual Afls

Overall scores: 6.3 out of 9 and 2.3 out of 3

This sample demonstrated the highest level of agreement between moderators and the actual Afls. What is particularly notable is not merely that moderators identified the correct inspection areas, but that many also identified the *specific developmental activity* implied by the Afls.

The clearest area of agreement centred on:

- improvement in RE teaching and learning
- assessment and questioning
- strengthening staff expertise through professional development
- improving the quality and consistency of prayer and liturgy

A significant proportion of moderator groups independently concluded that staff development was required. This is especially important because CPD is not only an inspection area in its own right but is also the principal means of improvement for addressing other weaknesses in a

school's provision. The fact that moderators inferred a need for training suggests the report contained repeated references to variation in practice, rather than isolated weaknesses.

Similarly, moderators did not simply identify 'collective worship' as an issue. Many identified:

- inconsistency between classes
- variability in staff confidence
- the need to disseminate best practice

These are much closer to the actual Afls than merely identifying a weakness in prayer and liturgy.

The report appears to have:

- identified a small number of recurring themes;
- revisited them in multiple sections;
- linked evaluation to impact;
- made the limitations sufficiently visible that readers could predict the improvement priorities.

In moderation terms, this is the strongest example of an Afl set that feels inevitable once the report has been read.

This sample suggests excellent coherence between the evidence gathered, the evaluation presented in the report and the resulting set of Afls. The moderators were largely reading the report in the same way as the lead inspector.

#### *Sample B – Some clear predictability, but only in one area*

Overall score: 3.63 out of 9 and 1 out of 3

Sample B presents an interesting picture because one Afl was predicted extremely well (hence the 1 out of 3 score for the second metric) whilst the others attracted much less consensus (thus bringing down the average score out of 9 to close to 3).

The overwhelming area of agreement concerned:

- questioning in RE
- challenge within lessons
- depth of pupil thinking

Many groups independently identified weaknesses around teacher questioning and opportunities for pupils to think more deeply. This indicates that the report narrative made this weakness highly visible. The correlation here was not merely thematic. Moderators often described the same kind of classroom activity that appeared in the actual Afl. This represents a high-quality match, perhaps the best single match of any of the Afls in the activity, but it was isolated in the sample. The other Afls generated much lower levels of agreement.

The actual Afls focused on:

- strengthening parish links;
- developing House Welcome Time and prayer provision.

Moderators occasionally identified these themes, but much less consistently. This suggests that references to these issues existed in the report but were not as dominant as the RE teaching issue.

The report appears to have contained one particularly salient weakness that drew attention away from others. Moderators were essentially saying: 'If we could only choose one thing this school should improve, it would be questioning in RE.' That indicates a report which strongly foregrounded one issue but perhaps did not establish the relative significance of the remaining Afls as clearly.

### *Sample C – Correct domains identified, but with weaker precision*

Overall scores: 3.25 out of 9 and 1.1 out of 3

Sample C is notable because moderators frequently identified the correct *areas* for improvement but less frequently identified the precise improvements eventually chosen by the lead inspector. On the face of it, the scores suggest greater predictability of Afls than Sample B, but the matches were much more general in character, and there were no predicted Afls that were a storing match to the actual Afls as was the case with Sample B. The scores reflect the general sense moderators had of where improvement was implied in the narrative but their guesses were much more widely dispersed than in either of the first two samples.

Moderators repeatedly identified:

- self-evaluation
- leadership monitoring
- prayer and liturgy development

This indicates that leadership capacity and evaluation processes were clearly visible within the report. There was considerable agreement that leaders needed sharper monitoring and stronger evaluation processes that were more rooted in the evidence. These are broadly consistent with the actual Afls.

Where moderators were less successful was identifying the specific emphasis on RE monitoring and evaluation. Many groups recognised a leadership issue but expressed it generically: 'Improve monitoring' rather than 'Improve monitoring of RE.' This suggests moderators saw the symptom evident in the report but not always the precise cause as the report author understood it.

The report successfully communicated that leadership improvement was required but was less successful at communicating exactly which leadership functions were weakest or which

improvements should take priority. In other words, moderators could see the destination but not always the route.

This was a reasonably coherent report, but the relationship between evidence and Afls is more diffuse than in Samples A or B.

#### *Sample D – Multiple plausible weaknesses competing for attention*

Overall score: 2.38 out of 9 and 0.5 out of 3.

This sample produced the lowest agreement and is therefore the most interesting from a moderation perspective. It is worth remembering that this report was for a school that gained grade 1s in almost every area, and this context probably goes a long way to explaining why moderators found it difficult to predict the Afls in the case of this sample.

Nevertheless, despite these difficulties, even here there were some recurring themes:

- pupil leadership of prayer;
- pupil participation;
- RE classroom practice;
- greater independence and ownership.

These appeared frequently across moderator responses.

The issue is that moderators consistently identified *different* weaknesses from those eventually chosen as Afls. For example: Moderators often suggested:

- improving pupil-led prayer;
- increasing creativity in liturgy;
- improving RE feedback;
- increasing independence in learning.

However, the actual Afls focused on:

- prayer reflections and requests;
- pupil participation in the Church's mission;
- sharing effective RE practice.

There is some overlap, but the connection is much weaker than in the other samples.

This report appears to have contained a larger number of minor weaknesses than the other samples. As a result, moderators were not drawn towards a common interpretation. Instead they dispersed across several different possibilities.

The key finding is not that moderators were wrong. Rather, many moderators made perfectly reasonable inferences from the text but arrived at different conclusions from the lead inspector. That could suggest a weaker connection between the narrative evaluation and the eventual Afls, but another explanation appears more plausible in the case of this outstanding school. In very strong schools both lead inspectors and moderators reading the report at a

later stage will struggle to find the marginal gains an outstanding school can make to become even better. The lack of consensus on this report is probably less a reflection of the quality of the report and more reflective of the difficulty of setting very strong schools meaningful targets for improvement.

## Conclusions

The moderation activity provides a useful indirect measure of report quality. A report with strong internal coherence should enable any reasonably astute reader to answer the question:

*'Having read this report, what are the three most important things this school needs to improve?'*

without seeing the Afls. On that basis the ranking of the samples provides a reasonable proxy for how successfully each report translated inspection evidence into a clear and convincing improvement agenda. What is particularly powerful about this exercise is that it moves beyond the usual question of *'Are these good Afls?'* and tests something more fundamental: Can well-informed readers and experienced moderators infer the Afls solely from the report narrative?

That is arguably a stronger test of report quality than many conventional moderation activities because a report can contain perfectly reasonable Afls, but if experienced readers cannot identify them from the evidence and evaluation presented in the report, then there is a potential weakness in the chain of reasoning. Conversely, where moderators independently converge on the same Afls, it provides strong evidence that the report has:

- correctly identified the most significant weaknesses;
- weighted those weaknesses appropriately;
- communicated them clearly;
- translated them coherently into improvement priorities.

One additional insight that emerged from the analysis is that the exercise appears to distinguish between *thematic agreement* and *operational agreement*. For example:

- Thematic agreement = moderators correctly identify the inspection area (e.g. Collective Worship, RE leadership, pupil participation).
- Operational agreement = moderators correctly identify the *actual improvement activity* required (e.g. 'develop pupil evaluation of prayer and liturgy' rather than simply 'improve collective worship').

The strongest reports were those where moderators were able to identify both with reasonable success.

This suggests three report-writing principles that could form the basis of future training on report writing:

### **1. The Afls should feel inevitable**

A reader should finish the report and think: 'Of course those are the Afls.' The best reports create a clear evidential path leading naturally to the improvement priorities.

## **2. Salience matters more than coverage**

Moderators were generally good at spotting weaknesses that were: repeated, developed, and linked to impact. They were much less likely to identify weaknesses that appeared only once or were described superficially.

## **3. Prioritisation must be visible**

Where several weaknesses appear equally important, moderators diverge. Where one or two weaknesses are clearly established as the most significant barriers to improvement, moderators converge. In many ways, the moderation exercise is therefore acting as a test of whether the report successfully communicates relative importance, not merely whether it identifies weaknesses.

Taken together these principles suggest that one key marker of quality for inspection reports is where the Areas for Improvement genuinely emerge from the report, rather than appearing to have been added afterwards, as an adjunct to the inspection evaluation rather than an integral part of it.

## Activity 3 – evaluating the quality of the focused sample

### Methodology

In the final task, moderators worked in the same cross-regional groups and were asked to provide comments on the quality of the sample reports, providing a summary written response in three areas:

- What went well (WWW)
- Even better if (EBI)
- Future training priorities suggested by the reports

These comments were collated using Microsoft forms and the common themes summarised for each sample.

### Findings

This third strand of the exercise is arguably the richest from a report quality assurance perspective because it moves away from judging outcomes (grades and Afls) and instead examines how experienced moderator readers respond to the report itself as a piece of evaluative writing. A particularly striking feature of the findings is that moderators demonstrated a high degree of consistency in their observations. In most cases, different groups identified the same strengths and weaknesses independently, suggesting that certain characteristics of the reports were readily apparent to experienced readers.

### *Sample A – grade 4 school*

There was a very high degree of consistency amongst moderators regarding Sample A. Most groups correctly identified the report as describing a school requiring significant improvement and generally agreed that the report successfully communicated the weaknesses that justified the judgement.

Moderators praised the report for its honesty and directness. The weaknesses of the school were clearly visible and most groups felt that the narrative provided sufficient evidence for readers to understand why the school had received such low grades. The relationship between narrative and judgement was particularly strong, which helps explain why moderators were highly successful in predicting the grades awarded.

However, discussion revealed significant concerns about the quality of the report itself. The most common criticism was that the report relied heavily on concepts that do not feature within the current framework, including differentiation, marking, feedback and provision for more able pupils. Moderators felt that the report occasionally judged the school against expectations that were not part of the CSI framework rather than against the framework itself. This concern was strengthened by the observation that the Areas for Improvement effectively amounted to three versions of the same recommendation: improve Religious Education.

The discussion therefore revealed a distinction between a report that clearly justified its grades and one that demonstrated strong framework fidelity. Moderators generally felt Sample A achieved the former more successfully than the latter.

#### *Sample B – grade 2 school*

Sample B generated the strongest consensus of all four reports. Moderators consistently praised the report for its quality of writing, evaluative tone and strong alignment between narrative and judgement. The report was widely regarded as a model of effective report writing. Participants particularly valued the clarity with which strengths and weaknesses were identified and the extent to which evaluative comments were rooted in evidence.

The discussion reinforced the view that the report demonstrated a strong understanding of the framework and that the grades awarded emerged naturally from the narrative. There was little disagreement regarding the quality of the report itself. Some moderators commented that the report lacked sufficient direct quotations from parents and pupils and an opportunity to celebrate the school's uniqueness had been missed, but this was a minority view.

The only significant, and widely-held criticism centred on one of the Areas for Improvement concerning parish engagement. Several moderators questioned whether this was a fair improvement priority because the quality of parish engagement is not always entirely within the school's control. This discussion highlighted a broader principle that while it may sometimes be legitimate to highlight parish partnerships as an improvement priority, Afls should focus on matters that schools can reasonably influence. On top of which, this Afl seemed particularly unjust in this instance given that parish relationships are actually praised as a strength of the school in the report itself.

Overall, however, Sample B attracted overwhelmingly positive feedback and was regarded as one of the strongest reports in the sample.

### *Sample C – grade 3 school*

Sample C generated the most critical feedback and produced the strongest consensus regarding weaknesses. While this report attracted more criticism than the other reports, moderators also identified a number of significant strengths. There was broad agreement that the report provided an honest and accurate account of the school's strengths and weaknesses and that the judgements awarded were generally well supported by the evidence presented. Moderators felt that the report did not shy away from difficult messages and was clear about the areas requiring improvement, whilst still recognising and celebrating the positive aspects of the school's work. Several groups commented that the report gave the school a realistic picture of its current position and provided a clear rationale for the grades awarded.

Moderators also noted that the report established a coherent relationship between the narrative and the eventual Areas for Improvement. Although concerns were raised about the wording of some AfIs, there was recognition that the underlying weaknesses identified by the report were genuine and that the improvement priorities broadly reflected the issues highlighted in the narrative. The report was therefore viewed as being fundamentally secure in its evaluative judgements, even if aspects of its expression could have been stronger.

A further strength was the report's level of detail. Moderators frequently commented that the report contained substantial evidence and provided a thorough account of the school's provision, leadership and outcomes. This enabled readers to understand the context of the judgements and contributed to the relatively high level of agreement amongst moderators when identifying the school's grades. Indeed, the moderation exercise demonstrated that most groups were able to recognise the report as describing a school operating around the Grade 3 threshold, suggesting that the narrative successfully communicated the school's strengths and limitations.

However, many felt that the report suffered from excessive advisory language. The narrative frequently told the school what it should do rather than evaluating what it currently did. As a result, the report often appeared to blur the distinction between evaluation and recommendation. This issue became even more apparent during discussion. Several moderators commented that the report contained so much advice that it was difficult to determine what was actually happening within the school. The report therefore became less evaluative and less precise than it might otherwise have been.

There was also a criticism of the reports failure to adequately capture typicality and too often referred to what happened on the inspection day itself, rather than using inspection evidence of indicators of what is typical. This was evidenced by its use of the past tense and by its justification of judgements with reference to what was seen on the day, rather than using these examples as one corner in the triangulation of evidence to establish what is typical of the school. Additional individual comments picked out the over-reliance on 'pupil-led' prayer as a

marker of quality in CW inspection area, which does not reflect the indicators of quality given in the framework itself for this area.

The Areas for Improvement attracted particular criticism. Moderators questioned whether the first Afl, which focused on self-evaluation across all judgement areas, was sufficiently specific. Although discussion ultimately confirmed that such an Afl is permissible, many moderators felt that it lacked the precision and clarity expected of high-quality improvement priorities.

Overall, moderators regarded Sample C as a report that was fundamentally accurate, honest and evidence-informed. Their criticisms were directed less at the inspection findings themselves and more at the style of report writing, particularly the tendency towards advisory language. As a result, the report was viewed as one that would benefit from refinement rather than one requiring substantial reconsideration of its judgements or conclusions. It appears to have been a fundamentally sound report whose principal weakness was stylistic.

#### *Sample D – grade 1 school*

Sample D was widely regarded as a high-quality report and generated strong agreement regarding its strengths. Moderators consistently praised the report's use of evidence, quotations and examples. Many felt that the inclusion of pupil, staff and parent voice gave the report authenticity and helped capture the distinctive character of the school. The report was frequently described as personalised, engaging and strongly rooted in the lived reality of the school.

The subsequent discussion provided important nuance. Moderators agreed that quotations are highly effective when used carefully, but cautioned that reports should not become collections of quotations. The strongest reports use quotations to illuminate wider findings rather than allowing individual opinions to replace evaluation. Several participants argued that quotations should represent wider views rather than isolated comments.

Some moderators found the report lacked balance in the religious education section, with some confusion between the outcomes and provision sections of the report in this area. This was confounded by several moderators commenting on the relative shortness of the report when dealing with outcomes and leadership in the religious education inspection area, with too much emphasis on provision. However, this was not a widely reported criticism and is perhaps more reflective of provision for religious education being the one area in which the lead inspector had most to say in terms of identifying weaknesses to be addresses.

The principal criticism of Sample D related to the nature of the Areas for Improvement. Moderators recognised that the school was operating at a very high level and that identifying meaningful improvement priorities was inherently difficult. The Afls were therefore described as “supererogatory” – improvements beyond what would normally be expected. While some moderators questioned the practicality of particular Afls, there was broad acceptance that

schools operating at this level require developmental recommendations that are necessarily more nuanced and aspirational.

Overall, Sample D was viewed as an example of strong report-writing practice, particularly in its effective use of evidence and stakeholder voice.

## Conclusions

The discussion revealed four recurring themes that cut across all four samples.

First, moderators repeatedly emphasised the importance of **evaluative writing**. Reports were most highly regarded when they clearly explained the impact of evidence and weakest when they drifted into description or advice.

Second, **advisory language** remains a significant concern. Moderators consistently identified indirect phrasing as one of the principal weaknesses of the report sample both in the morning and the afternoon. This issue was particularly evident in Sample C but was visible elsewhere as well.

Third, **Areas for Improvement** continue to be the least secure aspect of report writing. Moderators repeatedly questioned whether Afls were sufficiently specific, evidence-based and aligned with the report narrative.

Finally, moderators highlighted the importance of **framework fidelity**. Several discussions revealed concerns that inspectors sometimes judge schools against personal expectations, historic practices or familiar educational concepts rather than against the framework itself. This emerged most clearly in relation to Sample A and suggests that continued training on framework application remains important. Advisory language, problematic Afls, confusion about pupil independence, misuse of liturgical terminology, and references to marking or differentiation all appear to be symptoms of the same underlying challenge: ensuring that inspectors evaluate schools against the CSI framework itself, rather than against inherited assumptions from previous frameworks or wider educational practice.

The exercise suggests that, across all four reports, consistency between narrative and grade is strong. Moderators can accurately predict grades given in the majority of cases. This would suggest that consistency between narrative and grade is a strength of report writers and, when contrasted with Activity 2 (below) grade integrity is generally stronger than Afl integrity. In other words, moderators were more successful at answering: 'What grade would I award this report?' than: 'What Areas for Improvement would I write?' This is encouraging from a grading perspective because it indicates that the evaluative language of the reports generally aligns well with the judgements awarded. The greater variability in the Afl exercise suggests evidenced in the next activity suggests that identifying and prioritising the most significant improvements remains the more challenging aspect of report writing.

## 4. Other updates and wider discussions

The chief inspector also took the opportunity to update attendees on other ongoing work and to invite a broader discussion about ongoing training needs and policy refinements for the September 2026 handbooks.

### Summary of discussions

#### Report quality, framework implementation and quality assurance

A significant part of the discussion focused on the findings emerging from the moderation exercises and what they revealed about current inspection practice. Participants acknowledged that report quality continues to improve overall, with particularly strong progress evident in consistency of grading, compliance statements and report clarity. The introduction of Airtable and revised reporting systems was recognised as having reduced technical errors and improved consistency, particularly in relation to grade aggregation.

Despite these improvements, a number of recurring quality assurance themes emerged. Participants discussed the continued use of advisory language in reports, noting that indirect phrasing can obscure evaluative judgements and make it more difficult for schools to understand precisely what is not yet secure. There was broad agreement that inspectors should be encouraged to use clearer and more direct evaluative language.

A second area of discussion concerned leadership sections. Participants noted that leadership narratives sometimes drift into descriptions of activity rather than evaluations of impact. There was also discussion about the distinction between provision and leadership, with recognition that inspectors occasionally blur the boundary between the two.

The quality of Areas for Improvement remained a major theme. Participants discussed the tendency for some Afls to be insufficiently specific, insufficiently grounded in evidence or unnecessarily distributed across all judgement areas. There was strong agreement that Afls should arise directly from the report narrative and should focus only on the most significant priorities for development.

#### Terminology and framework understanding

A substantial discussion focused on terminology and interpretation of the framework. Participants identified continuing confusion regarding the use of terms such as collective worship, prayer, liturgy, assemblies and chaplaincy. Particular attention was given to ensuring that inspectors use terminology consistent with the Prayer and Liturgy Directory and current Church understanding. For example, the discussion clarified that 'assembly' refers to any large gathering, while 'celebration of the word' is a specific liturgical event, that may happen during an assembly, but the terms were not synonymous. In addition, it was discussed that 'liturgy' should be reserved for Mass and sacraments, while 'prayer and liturgy' can be used as a catch-all for various forms of worship in schools, it is usually more helpful to speak about prayer for

individual activities, or for celebrations of the word where that is what is actually being referred to. The importance of using these terms as understood by the Church was emphasised.

Related to this was an extended discussion concerning pupil leadership and independence within prayer and liturgy. Participants agreed that independence should not be treated as a measure of excellence in its own right. Rather, the quality of pupils' participation, planning, ministering and evaluation should remain the primary consideration. It was recognised that some inspectors and schools continue to interpret these aspects of the framework inconsistently.

Further discussion addressed Catholic Social Teaching. Participants noted a tendency for reports to place disproportionate emphasis on this area, sometimes at the expense of broader consideration of the Catholic curriculum and other descriptors within Catholic Life and Mission. There was agreement that further guidance would be beneficial once the Catholic Life and Mission Directory becomes available.

The discussion also touched on Relationships and Sex Education, where concern was expressed that reports occasionally omitted reference to RSE altogether or commented on aspects of quality that lie beyond the scope of inspection under the current framework.

#### Framework revision and future developments

Participants received an update on the national framework review. It was reported that almost 1,000 responses had been received to the consultation survey and that qualitative analysis of responses was underway. The findings will inform future revisions to the framework and are expected to form a major focus of future meetings with coordinators and inspector training events.

There was also an update on the work of the Catholic Multi-Academy Trust (MAT) Inspection Working Party. Discussion focused on the development of a framework capable of inspecting Catholic MATs without increasing the burden on individual schools. Participants emphasised the importance of maintaining school-level inspection whilst recognising the need to evaluate governance and leadership structures operating at trust level. It was suggested that leadership and governance may become a distinct framework category in future iterations of the framework.

An update was provided on the development of the Joint Inspection Framework. Participants were informed that the draft framework had been approved by the bishops, that the first phase of inspector training had been completed, and that further training and pilot inspections would take place during the next academic year before final implementation.

### Inspector development and future moderation arrangements

The discussion also explored future training priorities and moderation arrangements. Participants identified several recurring areas requiring further professional development, including:

- Areas for Improvement;
- evaluative writing;
- framework terminology and fidelity;
- leadership evaluation;
- distinctions between chaplaincy, prayer and liturgy.

There was support for strengthening expectations around attendance at lead inspector training and for revising the annual training cycle to ensure inspectors receive both core and role-specific development opportunities, using the model proposed:

- Autumn term: training for all inspectors
- Spring term: focused team to lead training in the North and South
- Summer term: in-person lead inspector training

It could be possible to introduce this approach from next year onwards, but it was also noted that the impending arrival of a revised framework for September 2027 might mean we need additional summer term training for all inspectors in Summer term 2027. A way forward was proposed of adopting this model in principle, always retaining the right to include additional meetings in case of need. It was also noted that the lead training would need to become an expectation for serving leads, rather than an option.

Participants also discussed future moderation arrangements. There was support for streamlining moderation activity into a single annual summer moderation event combining report moderation and quality assurance findings, whilst retaining an autumn meeting focused on strategic developments and framework matters. It was further suggested that moderation materials could be distributed in advance to allow participants additional preparation time and to allow for the reading for moderation part of the meeting to be completed in advance, leaving space for the discussion to take place during the meeting itself. This would allow the retention of both morning and afternoon aspects of the moderation day to be retained but in a streamlined way.

### Ofsted inspections and inspector eligibility

A further discussion concerned inspectors who receive multiple “needs attention” outcomes during Ofsted inspections. Participants recognised the need for a clearer and more transparent approach to determining whether such outcomes should affect an individual’s eligibility to inspect. It was agreed that further analysis of published Ofsted outcomes would be undertaken before any policy proposals are developed and that a proposal would be presented to the coordinators at their next summer term meeting.

### Early Years, Reception and the 10% Religious Education requirement

Participants discussed continuing uncertainty regarding the application of the Religious Education Directory requirement that at least 10% of curriculum time should be devoted to Religious Education in every year of compulsory schooling. Particular attention was given to Reception classes, where inspectors and schools have occasionally reached different interpretations of how this requirement should be evidenced within a continuous provision model.

It was agreed that the requirement clearly applies to Reception because Reception forms part of compulsory schooling, even though not all pupils attend full-time from the beginning of the academic year. However, participants recognised that identifying and evidencing Religious Education within a continuous provision environment can be more complex than in later phases of education. The discussion concluded that the issue is not whether the requirement applies, but how inspectors and schools should understand and evaluate its implementation within an Early Years context. It was noted that, in practice, Reception pupils may often receive significantly more than 10% Religious Education, but that this learning is frequently integrated across provision and therefore less readily quantifiable.

Participants reported instances where inspectors and schools had disagreed about the interpretation of this requirement, suggesting a need for greater clarity and consistency. It was therefore proposed that future editions of the Inspection Handbook should include explanatory guidance on how the 10% expectation should be interpreted and evidenced within Reception continuous provision.

### Action points from the day

- **Inspector Training priorities:**
  - Further training on Afis, focussing not just on quality but on coherence with the report narrative.
  - Further training on evaluative writing, particularly in relation to leadership sections of the report to help inspectors focus on impact.
  - Review and clarify the use of terminology such as "collective worship", "assembly", "liturgy", and "prayer" in inspection reports to ensure consistency and correct application in line with the framework and provide refresher training for inspectors on pupil leadership of prayer and liturgy, emphasising the correct interpretation and application in reports and AFIs.
  - Develop and disseminate guidance on the distinction between chaplaincy and prayer/liturgy in inspection reports to improve clarity and accuracy.
- **Catholic Social Teaching Coverage:** Review and balance the coverage of Catholic social teaching in reports, ensuring it is not overemphasised at the expense of other descriptors in the framework and particularly to support an understanding of what is meant by 'the Catholic curriculum'. Create and share examples and guidance to help

inspectors and schools better understand and evidence what constitutes a strong Catholic curriculum in practice. (the team)

- **AFI Appropriateness and Consistency:** Analyse post-training and pre-training AFIs to determine if recent Spring term training has improved the appropriateness and referencing of AFIs in inspection reports.
- **Inspector Feedback and Reporting:** Provide feedback to lead inspectors whose reports had been used in the sample, ensuring that feedback to inspectors following this meeting is balanced and not overly negative, including a caveat to acknowledge improvements and context.
- **Future QA and moderation meetings:** For 2026-2027 the QA and moderation meetings are to be consolidated and coordinators asked to carry out some pre-meting analysis of samples in cross-hub pairs or groups to facilitate more efficient use of the national moderation day.
- **Ofsted and inspector eligibility:** PR to provide an eligibility proposal for the summer term meeting of the coordinators to settle the relationship between the new Ofsted report card and ongoing inspector eligibility.
- **Future meting pattern:** PR to discuss with the advisory committee the overall reduction of inspector training expectations to follow the pattern of Autumn for all; Spring for team to lead; Summer for lead inspectors and to publish updated CSI calendars for 2026-2027 onwards.
- **Curriculum %age clarification for Reception class:** The advisory committee to review the handbook section on compliance and add explanatory guidance to the handbook concerning the application of the 10% RE requirement in Reception and continuous provision contexts.

## 5. Overall conclusion

Taken together, the findings of this report present a strongly encouraging picture of the current quality of inspection reporting across the Catholic Schools Inspectorate, while also identifying a focused set of priorities for further development. Across both the morning and afternoon moderation exercises, moderators found clear evidence that inspection reports are becoming more accurate, more consistent and more professionally assured. In technical terms, most reports now meet report-writing requirements securely. Aggregation errors are rare, compliance statements are generally stronger, narrative and grades are usually well aligned, and the overall quality of evaluative writing has improved when compared with previous years. In both the quantitative and qualitative moderation data, the 2025–2026 sample performed more positively than the equivalent samples from 2024 and 2025 in almost every area measured. This provides strong evidence that national moderation, quality assurance processes and inspector training are continuing to have a positive impact.

At the same time, the report makes clear that improvement remains necessary in a relatively small number of recurring areas. Most notably, Areas for Improvement continue to be the least secure element of report writing. While inspectors are generally successful in communicating grades clearly through the report narrative, the relationship between narrative and Afls is often less secure. Moderators could usually infer the grades awarded from the text with a high degree of accuracy, but they were much less likely to identify the same Afls as the lead inspector. This indicates that, although report writers are increasingly effective at explaining how schools are performing, they are not yet consistently as effective at showing why the chosen improvement priorities are the most important or how they emerge from the evidence presented. The findings suggest that the next phase of improvement in report writing should focus less on grade integrity and more on Afl coherence, specificity and precision.

A second major conclusion concerns the quality of evaluative writing. Moderators repeatedly affirmed that reports are strongest when they explain impact clearly, distinguish evaluation from description, and avoid slipping into advisory language. Where reports became less effective, this was often because they described activity rather than evaluating its significance, particularly in leadership sections, or because they used indirect recommendations rather than direct evaluative judgements. Closely related to this is the question of framework fidelity. Several of the issues identified during the day, including advisory language, confusion around pupil independence, uncertainty over prayer and liturgy terminology, and occasional references to concepts not found in the framework, all point to the same underlying challenge: ensuring that inspectors judge schools against the CSI framework itself rather than against inherited assumptions or wider educational habits. This is not a systemic failure, but it remains an important developmental priority.

The wider discussions and action points recorded at the end of the day show that the inspectorate is well placed to respond to these findings. There is now a clear and evidence-based agenda for future work: further training on Afls, evaluative writing and framework terminology; continued refinement of moderation and quality assurance processes; clearer national guidance in areas of recurring ambiguity; and an increasing expectation that moderation evidence should inform professional development in a focused and proportionate way. Overall, this report does not point to weaknesses in the inspection system, but to an inspectorate that is maturing in confidence, consistency and self-evaluation. Report quality is clearly stronger than it was two years ago, moderators are reading with greater sophistication, and the remaining challenges are now more precise and therefore more manageable. The overall conclusion is therefore positive and constructive: national report-writing practice is improving steadily, the quality assurance framework is working, and the priorities identified through this moderation exercise provide a clear basis for the next stage of inspector development and national consistency.