

Name	Joseph Martin
Home Institution	Durham University
Name(s) of course(s) examined <i>e.g. Tripos Part/ MPhil/ MRes</i>	NST Part II History and Philosophy of Science
Academic year of examination	2024-25
Level (<i>Delete as appropriate</i>)	Undergraduate
Year of Appointment	1 st

	Yes	No	N/A
1. Are the academic standards set for the award appropriate for the qualification, and comparable with similar programmes in other UK institutions?	√		
2. Are you satisfied that you received sufficient programme materials (handbooks, regulations, marking and classing criteria) in a timely manner?	√		
3. Are you satisfied that you were consulted adequately on draft examination papers, and that your comments and suggestions were taken into consideration?	√		
4. Are you satisfied that the assessment was pitched at the appropriate level?	√		
5. Was the general standard and consistency of marking appropriate?	√		
6. Do the assessment processes measure student achievement rigorously and fairly against the intended outcomes of the programme?	√		
7. Are you satisfied that issues raised on your previous report form have been properly considered and, where applicable, acted upon?			√
8. Did you receive a written response from the Department to your previous report form?			√

If you replied No to any of the questions above, please expand here:

N/A

Do you have any concerns about the course, including standards and quality?

No

Are you satisfied that the procedures associated with the assessment are efficient (e.g. timeframes, draft papers, questions, design and conduct of exam, meetings, vivas)?

Yes

Do you have any comments on marking and classing (e.g. range of marks, action around borderline marks, penalties, moderation, double marking, reconciliation of marks)?

Range and Spread of Marks

The spread of marks and deployment of the scale were appropriate. Double marking typically drives regression toward the mean, but this cohort wasn't particularly clumpy. Unsound work was marked down appropriately; excellent work was duly rewarded. Extreme marks were rare, but not due to unwonted conservatism. Though few scores reached the high 70s or low 80s, the criteria for that range set a lofty bar and it stands to reason that, on a small course, relatively few pieces of work will peek over it. The best work was in no way sold short by not having been marked higher.

Process

The examining process is robust and transparent. In borderline cases, and where discrepancies straddled boundaries, the department employs sound reconciliation procedures. The evidence I was provided clarified the reasoning that informed the agreed marks. It is notable that *all* work is double marked, ensuring a high standard of oversight.

Patterns meriting attention

Mean exam marks (HPS and BBS) were comparable for history and philosophy. But on coursework assessments, philosophy's mean (67.94) was appreciably lower than history's (70.13). The effect remains (slightly amplified) when excluding the top and bottom two marks from each set (67.63; 70.64). Given the small-n student body, this could be an artefact of student choice, but I nevertheless encourage monitoring the difference over coming years. The department has helpfully resolved to collate data since 2021-22 to investigate the extent to which the effect is persistent. In any event, it will be useful for HPS examiners to compare notes and ensure that the history and philosophy scales remain in harmony.

Examiners queried whether mark variance differed between disciplinary groups. The variance of marks showed only a very modest difference. History coursework marks differed from the mean by 5.45 marks on average, philosophy by 5.18, but this small difference might have been erased or reversed with a different performance by a single student. The average variance of exam marks in HPS show no meaningful difference (4.51 history vs. 4.54 philosophy). The gap was larger for BBS exams (4.50 history vs. 5.89 philosophy), closes when excluding several very low marks on the BBS philosophy exams. In other words, history and philosophy markers appear to be availing themselves of a similar spread of marks.

Both BBS major subject philosophy papers (NST2BBS 13_3 and NST2BBS 13_4) had large average discrepancies in first and second markers' scores (5.5 and 6.5 respectively). For comparison, two HPS exam papers also had average discrepancies above 5, but this resulted from an outlier on one extreme of the mark scale and/or a small sample size. In these BBS exams, in contrast, it resulted from high discrepancies on multiple scripts. The process for resolving these discrepancies is robust and the consensus marks are consistently apt, but I'd encourage ongoing discussion among examiners of the standards and expectations for BBS major subject philosophy exams.

Do you have any comments on the student experience of the course and/or their experience of the assessment process?

This cohort performed overwhelmingly better on coursework than on exams. Just six of the twenty-eight students scored higher on the exams (two of those by less than one percentage point). The remaining twenty-two performed better on coursework, all but one of whom by more than two points and sixteen of whom by more than five. The quality and originality of the primary source essays and dissertations I reviewed reinforces the impression that students gain a great deal by engaging in these assessments.

Do you have any comments on University policies (e.g. the role of the external examiner, policies around plagiarism, script annotation)?

As discussed in the examiners' meeting, it would be useful to clarify policy around the use of supervision essays (or other formative work), in whole or in part, in assessed work. I make no recommendation about what the policy should be, but unambiguous guidance would be helpful for both students and examiners.

Please describe here any recommendations for improvement.

Examiners flagged the three-hour open-note exam as an area for improvement. I concur. Experimenting with exam delivery is a sensible response to disruptions following Covid and challenges posed by LLMs. It is not clear, however, whether this year's format was optimally successful in achieving its pedagogical or evaluative aims.

- The pedagogical question: Open-note exams are a reasonable response to the difficulty of effectively policing closed-note take-home exams, but they incentivise neither the broad knowledge acquisition

that sat exams usually aim to promote nor the deep analysis valued in coursework assessments. If the department persists with this format, it would be useful to articulate clearly the pedagogical goals it might achieve more effectively than sat exams or coursework, if any, particularly in light of...

- The evaluative question: Examiners hold the well-founded conviction that at least some, if not most students will employ LLMs in exams—in various ways, spanning the integrity gradient. Because of the constraints on the analytical depth students can be reasonably expected to generate in three hours, LLMs can imitate exam answers more successfully than they can imitate coursework. Consequent challenges include markers' struggles to set aside suspicions when they encounter error-free prose and the fact that even flagrant misconduct is prohibitively difficult to evidence.

The simplest short-term response to these challenges would be a return to sat, closed-note exams. That format would restore the incentive for students to study a broad range of material and obviate concerns about LLMs.

Looking to the horizon, though, challenges with the exam format represent an opportunity to reflect on how assessment can and should adapt to a changing technological context. I would encourage discussions about ways to diversify forms of assessment and deemphasise those forms most susceptible to LLM use.

In particular, I encourage exploring iterative assessment, in which students revise and resubmit work based on feedback. Beyond educational contexts, most writing that matters is *rewriting*, so such assessments better reflect the craft of HPS scholarship. Further, humanities programmes must proactively envision how students' skills will change over the next 3–5 years, as we meet student who have had access to LLMs over progressively greater proportions of their intellectual development. Teachers of the humanities are likely to have to focus more explicitly on the processes of research and writing, which might include finding ways of assessing *how* students write alongside assessing the results of that process. Iterative assessment can be an effective tool for doing so, creating opportunities to develop skills students might have had limited opportunity to acquire, sharpening their critical judgment, and reducing the incentive to use LLMs irresponsibly.

Finally, the department might consider in a systematic way if, how, and where AI literacy should be included in the programme structure. This is apt to be particularly useful for students coming from the natural sciences, where attitudes toward AI tools often differ from attitudes in the humanities.

All of the above should, of course, be approached with workload considerations squarely in view, and while retaining a commitment to imbuing students with a breadth of knowledge, alongside depth.

Please highlight any good practice you encountered.

The dissertation and primary source essay assignments inspired an excellent body of work by encouraging detailed, in-depth engagement with materials and concepts central to the field.

As noted above, marking and reconciliation procedures in the department are exemplary, as is the conscientiousness of examiners' engagement with assessment processes.

Have you seen any evidence of grade inflation?

No. Eight students earned first-class degrees in HPS Part II. This represents approximately 30% of the cohort—broadly consistent with previous years, in which about 20–50% of students have earned firsts. In a programme with relatively small and variable student numbers, the absence of year-to-year fluctuations would be curious. I see no indication that anticipated fluctuations are orchestrated by marking practices.

The examiners were constructively reflective about the sort of factors that contribute to fluctuations and discrepancies, both with respect to degree classifications and with respect to performances on individual papers. In the latter case especially, differences in performance are to be expected based on the different preparation students enrolled in different departments bring. I was thus satisfied that the examiners are engaged in the process in a way that makes grade inflation unlikely.

If this is your final year as external examiner? If so, have you seen improvements over your tenure? Has the Department acted on your advice?

N/A

Do you have any other comments?

No