Judging People, Evaluating Work, or Encouraging Virtue

Antony Eagle antonyeagle.org

2023-10-31 » Festival of Learning and Teaching

The Standard View: Assessment *of* Learning

Assessment Measures Student Attainment

- *Higher Education Standards Framework (Threshold Standards) 2021*:

> 3. Methods of assessment are consistent with the learning outcomes being assessed, are capable of confirming that all specified learning outcomes are achieved and that grades awarded reflect the level of student attainment.
> 4. On completion of a course of study, students have demonstrated the learning outcomes specified for the course of study, whether assessed at unit level, course level, or in combination. (Department of Education and Employment 2021: §1.4)

  - The QAA in the UK says something similar: assessment 'determines whether each student has achieved their course's learning outcomes and allows the awarding body to ensure that appropriate standards are being applied rigorously' (Quality Assurance Agency for Higher Education 2018: 2).

- Assessment is for **checking and evaluating attainment**; it is for 'assuring learning outcomes' and 'forming trustworthy judgements about student learning' (Lodge, Howard, *et al.* 2023: 1–2).

An Problematic Ambiguity

- There are real concerns about **implementation**: about whether our approaches are fair and transparent.
- But I have some deeper concerns about the standard approach, even done well. It is, I think, **ambiguous** between two senses of *student attainment*.

  Direct Assessment is about judging how well students *achieve* the Learning Outcomes and embody Graduate Attributes.

  Product Assessment is about evaluating how Learning Outcomes and Graduate Attributes are *displayed* in submitted work.

- The problem is that while we are required, legislatively, to aim at Direct assessment, what we do in practice is Product assessment.
  - Product assessment itself is a mere means to an end; the solution obviously is not to begin thinking of our role as determining the quality of outputs independently of how they are produced.

Disguising the Gap

- To disguise this gap, we are asked to make learning outcomes attributes of submitted work, rather than attributes of students.
- This is reflected in 'SMART' approaches to the design of learning outcomes. We are enjoined to make LOs **measurable**, where that amounts to the positivist idea that attainment of the LOs must *essentially manifest in behaviour*.
  - 'Students will **demonstrate familiarity with** the basic ideas of field *X*' is preferred to 'Students will **understand** the basic ideas of field *X*'.
- Of course even this doesn't do the job; student work might not reflect their attainment of the LOs, even when the LOs are measurable.
  1. Work might be better than the student's level of attainment, whether by design (integrity violations, which will be our focus below) or accident (a charitable marker sees a poorly expressed misunderstanding as a subtly novel interpretation of a text).
  2. Student attainment might be better than the work (a student who through nerves or circumstance fails to show in their work a skill they in fact have: Radford (1966))

Graduate Attributes

- No such disguise is available for the Graduate Attributes. E.g.,

> Graduates have comprehensive knowledge and understanding of their subject area, the ability to engage with different traditions of thought, and the ability to apply their knowledge in practice including in multi-disciplinary or multi-professional contexts. (Adelaide Graduate Attributes)

- Here there is no behaviourist emphasis on manifestation – the GAs reflect what we **really value** in our graduates:
  - **Knowledge** – not the regurgitation of facts under exam conditions;
  - **Understanding** – not the appearance of it in an essay;
  - **Ability** – not the demonstration of that ability on a specific occasion.
- **Why should we believe that tasks that build knowledge and ability are those that permit straightforward measurement of achievement?**

Integrity

- The gap between what we are supposed to assess, and what we have access to, is emphasised by **breaches of academic integrity**. Seeing assessment as concerning the certification of student attainment requires a strong and reliable **link** between the student and the piece of work submitted for assessment.
- I understand such breaches broadly: they are cases where a student submits, as evidence of their attainment of LOs, material which is in fact intentionally misleading about their attainment (to the extent that LOs are appropriately reflected in assessment tasks).
  - Intended to cover traditional plagiarism, contract cheating and generative AI, the use of automated translation tools; but rules out mistakes, etc.
  - A salient reminder that not **responding appropriately** to some evidence – e.g., by being an 'easy' grader – is also a violation of academic integrity!
- The work serves as a proxy for the student; its qualities are taken to reflect the student's level of achievement and evidence of their embodiment of the graduate attributes.
- If our primary purpose is certification of attainment, we must impose robust and effective **deterrents** for violations of academic integrity.
  - Note: not because it is **wrong**, or at variance with **academic values** – but because the purpose of the assessment regime is predicated on it.

The Threat of AI

- This has always been a weakness of this conception of assessment: it has always been possible for students to mislead about their ability, submitting work that is not their own (either cheaply, by plagiarising, or expensively, by engaging in contract cheating).
- What's relatively new is the possibility of cheap contract cheating, using **generative AI** such as ChatGPT – or the superior Claude.
  - Let's not get carried away: these models are expensive to run and it is not likely that the current phase of lots of free usage will continue indefinitely – even a small fee, such as the USD 20 per month for the 'Pro' version of Claude, will deter many students.
- But there is a challenge here: what should we do if it is likely that some substantial fraction of our students are breaching academic integrity?

Hostile Assessment

Design Out the Risk: Hostile Assessment

- *Hostile architecture* is the term for aggressively unfriendly street furniture designed to deter loitering and occupation by the unhoused (e.g., spikes to derail skateboarders, benches too narrow to sleep on, etc.)
- I introduce **hostile assessment** to label attempts to design assessment tasks that make unwelcome behaviours – breaches of academic integrity – so unpleasant or difficult that students won't or can't breach.
    - This is the path chosen by those who think the right answer to generative AI is to return to 3-hour in person exams, and other completely inauthentic but hard-to-game forms of assessment.
    - This is, unfortunately, part of what is provisionally recommended by TEQSA (Lodge, Howard, *et al.* 2023): the proposed model is to adopt a holistic model of assessment at the program level (good!), safeguarded by the widespread use of hostile assessment: orals, observed clinical tasks, practicals, interactive code reviews, and timed exams (bad!).

Problems with this approach: here are some…

Workload Programmatic assessment is great in theory, but requires evaluation of the connections between assignments, in addition to marking each individual assignment. Automatic assessment of MCQs in Canvas is a boon to staff, but problematic from an integrity perspective.

Unfairness/Nontransparency Most hostile assessment – apart from exams – is hard to anonymise.

‘Holistic’ forms of assessment are especially prone to bias and nontransparency – it is built in, for example, that you should base your judgments on *this* piece of work in part on the student’s track record!

It is hard to design plausible and reliable criterion for programmatic assessment.

… and some more

Authenticity Hostile assessment – like hostile architecture! – is generally **less functional**. * It is artificially constructed and less **authentic** to the discipline or to the realities of the workplace.

Responsibility for Learning Good assessment regimes should not involve **forcing** students to do anything, or giving them no option. * Why? Because someone who does something only because they could not do otherwise is not morally responsible for doing it (Frankfurt 1969: 839). * If we want our students to be **ultimately responsible for their own learning**, then we need to give them autonomy in approaching it, even if that gives them the autonomy to do the wrong thing. * Integrity isn't about doing the right thing only because you want to avoid being punished.

The Big Problem

Misalignment Choice of assessment tasks is driven by concerns of integrity rather than the needs of the course.
Why should we believe that tasks that build knowledge and ability will also permit reliable defences against cheating?

- Learning outcomes themselves are chosen for measurability and ability to force compliance with integrity codes, rather than chosen to reflect desired learning.
- This is exactly the wrong way around: good course and program design puts learning first: we ask, *what do we want students to learn?* and then ask *what tasks and activities, assessed or not, will support that learning?*

Assessment *For* Learning

A Better Way: Put Learning First!

- The purpose of assessment is to **facilitate learning**, not to evaluate it retrospectively. (Sometimes termed *assessment for learning*.)
- The institution wants graduates to emerge in possession of the graduate attributes. The design of courses, and assessments within courses, should begin by asking: how can we encourage students to cultivate these attributes?
  - Note almost none of them are directly measurable in behaviour, though it might be a defeasible guide to them: knowledge, ability, understanding, ethical competency, self-awareness…
- The point of assessment, on this view, is to give students an **opportunity to cultivate** the graduate attributes – to become virtuous, knowledgeable, responsible citizens.
- The point of assessment isn't to check student attainment, but to **encourage** it.
  - Of course checking attainment is still a **subsidiary** aim of assessment – even exploitable assessment regimes can be reliable measures of attainment (so long as *reliability* doesn't amount to a guarantee).

Defending the Essay

- The bluntest version would say this about some recently derided forms of assessment:

  > If (i) the essay is the best way for students to foster and cultivate the skills of constructing coherent long-form arguments, integrating and weighing up a diversity of evidence and (ii) those skills are what you want students to gain from taking your course, then (iii) you ought to set an essay as a principal assessment task – even if it cannot be proofed against generative AI.

  - Some will say this is no longer a skill we need, especially in the presence of generative AI that will do all our longform writing for us.
  - On the contrary: in the presence of vast quantities of generative AI bullshit (Frankfurt 1986), we will need those skills of sifting competing evidence and thinking things through ourselves more than ever.

Managing Academic Integrity

- The long-noticed tension between assessment **of** learning and assessment **for** learning crystallizes here (Norton 2007: 134).
- It simply may not be possible to preclude breaches of academic integrity – just as it is not possible to stamp out other forms of contract cheating (Dawson 2022).
- The key is to **incentivise good behaviour**. Students cheat for lots of reasons (Brimble 2016; Amigud and Lancaster 2019); and if we can shift the amount students are willing to invest we can drive compliance without it being our primary aim.
  - There is no need for hostile architecture if the drivers of so-called ‘anti-social behaviour’ are addressed.

Assessment Design

- Good assessment is **intrinsically motivating**: if students (i) know what skills/knowledge the task is intended to support them in developing and (ii) believe that those skills/knowledge are going to be valuable to them in the future, they have an interest in doing the task.
- Students are motivated to **avoid failure** in a situation of **time-scarcity** (Brimble 2016: 375); probably a more important driver of integrity breaches – with respect to use of generative AI – than a desire for high marks.
  - Assessment regimes that remove the stress of failure – such as **specifications grading** (Nilson and Stanny 2014), which empowers students to choose their own desired level of attainment, increasing autonomy – might be of great benefit.

Culture Change

- Emphasise that academic integrity is a broader ethical enterprise, not merely about 'not plagiarising'. Academic integrity is

  > the expectation that teachers, students, researchers and all members of the academic community act with: honesty, trust, fairness, respect and responsibility' (https://www.teqsa.gov.au/students/understanding-academic-integrity/what-academic-integrity)

- As such acting in accordance with academic integrity is part of a broader commitment to morality.
- There may be room for behavioural **nudges** here.
  - Maybe de-emphasise the stakes and the credentialing role of assessment – this encourages a **transactional** approach to assessment rather than an ethical one,
  - Maybe get students to sign an 'honour code' on submission, an overt promise that the work is their own.
    - That said, the key work on whether signing at the start 'makes ethics salient' and decreases dishonesty (Shu, Mazar, *et al.* 2012) now been retracted because 'two different [authors] independently faked data for two different studies' reported in the paper (Simonsohn, Nelson, and Simmons 2023)!

Teacher, Bureaucrat, Cop

> Clearly, some students will start to cheat using automated paper-writing. People are proposing responses that involve going full surveillance state. They have proposed making students write their papers with eye-tracking software on, or writing all papers in controlled and secure environments, like testing centers. These choices serve the functions of the bureaucrat and the cop—but not, I think, the teacher's. Because the environment they are suggesting—an environment of surveillance, paranoia, and profound distrust—is deeply hostile to some of our subtler educational goals. It is hard to turn in a creative expression, to really reflect on your values and world-view, if you have to write your essay in a single session in a monitored testing center under a camera's baleful eye. Going full surveillance may catch some cheaters, but at the expense of providing a richer, more supportive educational environment for the rest of our students. (Nguyen 2022)

Appendix: Implementation

Implementation: Transparency and Fairness

- To start, let's ask: *do our current assessments fulfil these goals?* Consider two issues.

  Fairness Are potential sources of bias controlled for? Would the same piece of work receive the same grade from different markers ? Would the same marker assign the same piece of work the same grade on different occasions (Price, Jhangiani, and Chiang 2015: 96)? Do assessment tasks measure what they are intended to (Price, Jhangiani, and Chiang 2015: 99)? Are assessment tasks inclusive, e.g., 'the assessment is non-threatening and non-anxiety provoking' (Rust 2007: 230)?

  Transparency Are the grounds for a given grade transparent to students being assessed? Can students perceive that assessment is fair and reasonable?

- Are we confident that our practice – and that of our colleagues – displays these desirable features (Rust 2007; Norton 2009: 143–44)?

Good practice

- Is everyone using a **rubric**, and a rubric that transparently links satisfaction of criteria to numerical grades assigned?
- **Anonymised grading**? It is harder to treat someone unfairly because of their group membership if you don’t know to which groups they belong (Dorsey and Colliver 1995).
  - Anonymous grading likely improves validity by eliminating potential confounders.
  - Anonymous grading can improve student **perception of fairness** too (Steele 1997; Falchikov and Goldfinch 2000); it sends a signal that demonstrated achivement is what matters (Carrington 1992: 565). (Though things are more complex when it comes to feedback and relationships: Pitt and Winstone (2018)).
- **Explicit reference to common criteria** in assigning grades – even just the generic grade descriptors?
  - Active training of students about how to engage with the criteria (Norton, Harrington, *et al.* 2005)?

References

References

Amigud, Alexander and Thomas Lancaster (2019) '246 Reasons to Cheat: An Analysis of Students' Reasons for Seeking to Outsource Academic Work', *Computers &Amp; Education* **134**: 98–107. doi:10.1016/j.compedu.2019.01.017.

Brimble, Mark (2016) 'Why Students Cheat: An Exploration of the Motivators of Student Academic Dishonesty in Higher Education', in T Bretag, ed., *Handbook of Academic Integrity*: 365–82. Springer Singapore.

Carrington, Paul D (1992) 'One Law: The Role of Legal Education in the Opening of the Legal Profession Since 1776', *Florida Law Review* **44**: 501–603.

Dawson, Phillip (2022) 'The Prevention of Contract Cheating in an Online Environment'. https://www.teqsa.gov.au/sites/default/files/2022-10/prevention-contract-cheating-in-online-environment-web.pdf?v=1587691121.

References (cont.)

Department of Education, Skills and Employment (2021) 'Higher Education Standards Framework (Threshold Standards) 2021'.

Dorsey, J K and J A Colliver (1995) 'Effect of Anonymous Test Grading on Passing Rates as Related to Gender and Race', *Academic Medicine* **70**: 321–23. doi:10.1097/00001888-199504000-00017.

Falchikov, N and J Goldfinch (2000) 'Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks', *Review of Educational Research* **70**: 287–322.

Frankfurt, Harry G (1969) 'Alternate Possibilities and Moral Responsibility', *Journal of Philosophy* **66**: 829–39.

References (cont.)

Frankfurt, Harry G (1986) ‘On Bullshit’, *Raritan* **6**: 81–100.

Lodge, Jason M, Sarah Howard, Margaret Bearman, and Phillip Dawson (2023) ‘Assessment Reform for the Age of Artificial Intelligence’. https://www.teqsa.gov.au/guides-resources/resources/corporate-publications/assessment-reform-age-artificial-intelligence.

Nguyen, C Thi (2022) ‘Teacher, Bureaucrat, Cop’.

Nilson, Linda B and Claudia J Stanny (2014) *Specifications Grading*. Routledge.

Norton, John D (2007) ‘Causation as Folk Science’, in Huw Price and Richard Corry, eds., *Causation, Physics, and the Constitution of Reality*: 11–44. Oxford University Press.

References (cont.)

Norton, Lin (2009) ‘Assessing Student Learning’, in Heather Fry, Steve Ketteridge and Stephanie Marshall, eds., *A Handbook for Teaching and Learning in Higher Education: Enhancing Academic Practice*: 132–49. Routledge.

Norton, Lin, Katherine Harrington, James Elander, Sandra Sinfield, Jo Lusher, *et al.* (2005) ‘Supporting Students to Improve Their Essay Writing Through Assessment Criteria Focused Workshops’, in C Rust, ed., *Improving Student Learning: Diversity and Inclusivity*: 159–74. Oxford Centre for Staff; Learning Development. https://www.researchgate.net/publication/313407459_Supporting_students_to_improve_their_essay_writing_through_assessment_criteria_focused_workshops.

Pitt, Edd and Naomi Winstone (2018) ‘The Impact of Anonymous Marking on Students’ Perceptions of Fairness, Feedback and Relationships with Lecturers’, *Assessment &Amp; Evaluation in Higher Education* **43**: 1183–93. doi:10.1080/02602938.2018.1437594.

References (cont.)

Price, Paul C, Rajiv Jhangiani, and I-Chant A Chiang (2015) *Research Methods in Psychology*. BCcampus. https://opentextbc.ca/researchmethods/.

Quality Assurance Agency for Higher Education (2018) 'UK Quality Code for Higher Education, Advice and Guidance: Assessment'. https://www.qaa.ac.uk//en/the-quality-code/advice-and-guidance/assessment.

Radford, Colin (1966) 'Knowledge: By Examples', *Analysis* **27**: 1. doi:10.2307/3326979.

Rust, Chris (2007) 'Towards a Scholarship of Assessment', *Assessment & Evaluation in Higher Education* **32**: 229–37. doi:10.1080/02602930600805192.

References (cont.)

Shu, Lisa L, Nina Mazar, Francesca Gino, Dan Ariely, and Max H Bazerman (2012) ‘Signing at the Beginning Makes Ethics Salient and Decreases Dishonest Self-Reports in Comparison to Signing at the End’, *Proceedings of the National Academy of Sciences* **109**: 15197–200. doi:10.1073/pnas.1209746109.

Simonsohn, Uri, Leif Nelson, and Joe Simmons (2023) ‘Data Falsificada (Part 1): ”Clusterfake”’.

Steele, Claude M (1997) ‘A Threat in the Air: How Stereotypes Shape Intellectual Identity and Performance’, *American Psychologist* **52**: 613–29. doi:10.1037/0003-066x.52.6.613.