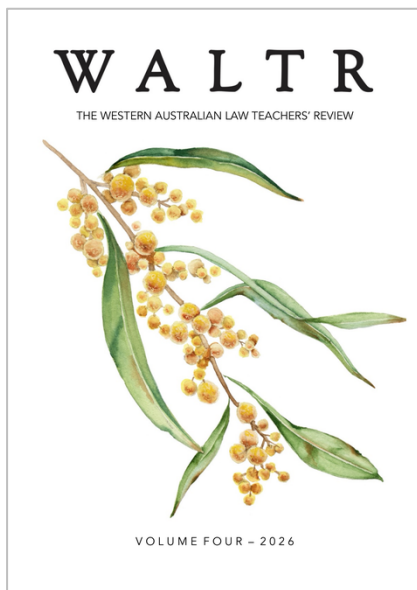




Western Australian Law Teachers' Review

Volume Four

ISSN: 2653-7710

The *Western Australian Law Teachers' Review* (WALTR) is a peer-reviewed scholarly journal which publishes research relating to legal education and related topics. WALTR is an independent, open access publication.

Journal homepage: www.WALTReview.com

Teaching Note

Roughly 68-ish, Give or Take: An Honest Assessment of Law School Exam Design

Julie Falck

Citation: Julie Falck, 'Roughly 68-ish, Give or Take: An Honest Assessment of Law School Exam Design' (2026) 4 *Western Australian Law Teachers' Review* 85.

Published: October 2026

Link to the volume containing this article: www.WALTReview.com/volfour

All articles in WALTR are published and made available at no cost and may be downloaded and used in accordance with the terms of the Attribution-NonCommercial-NoDerivatives Creative Commons Licence — CC-BY-NC-ND 4.0. In summary, you must give appropriate credit to WALTR and the author, you may not use the material for commercial purposes, and you may not distribute any derivative material.

ROUGHLY 68-ISH, GIVE OR TAKE

AN HONEST ASSESSMENT OF LAW SCHOOL EXAM DESIGN

JULIE FALCK^{*}

This piece is published as a 'teaching note' rather than a scholarly article.

WALTR publishes teaching notes to provide space for reflective, practice-based contributions that sit outside the scope of traditional peer-reviewed articles but are nonetheless of value to the community of legal educators. They may include reflections on teaching and assessment, contextualised lesson plans (including for secondary legal studies), or opinion and commentary on broader issues in legal education.

I Critically analyse the proposition that a mark of 68 is a reliable indicator of a student's actual capability. (10 MARKS)

A mark is a small thing that makes some surprisingly large promises. When we hand back a script marked 68, we are claiming — whether we would put it this way or not — that the number reflects what the student can actually do; that another marker, on another afternoon, would have landed somewhere close; and that the student had a fair shot at earning it. The assessment literature calls these validity, reliability and fairness. Most of us signed up to teach law, not to design measurement instruments, and the conditions under which we set exam papers rarely allow for unhurried reflection. Yet exams decide the largest share of most students' grades.

Nobody taught me how to write an exam, and I suspect nobody taught you either. The standard induction is inheritance: you read the unit's past papers, absorb the house style, and produce another one like them. Past papers are a template, and a template is a wonderfully efficient answer to the wrong question: it tells you what an exam here looks like, and nothing about what your exam should do. So I set myself a thought experiment.

^{*} Julie Falck is a Lecturer at the UWA Law School, The University of Western Australia.

Suppose you had to design a law school exam from scratch — from the literature and the evidence, inside the real institutional rules, but owing nothing to the filing cabinet. How different would it look from the exams we actually set? Picture every exam as a position in a triangle whose corners are validity, reliability and fairness, drawn inside a frame of feasibility: marking time, institutional rules, curriculum, cohort size. There is no ideal point; intentional design means choosing your position and knowing what it costs, rather than letting the frame choose for you. My answer, after a long while in the scholarship and a systematic trawl through our own past papers, is: not terribly different, but different in unexpected places. The inherited exam turns out to be broadly defensible in its architecture, but strangely careless in its small print.

II 'Assessing complex achievement is judgement, not measurement.' Discuss with reference to validity, reliability, and fairness. (20 MARKS)

Validity is an inference, not a property. The most useful idea I found is Peggy Maki's reframe: validity is not something an exam *has*; it is a question of how well the task prompts students to show the learning we care about.¹ Michael Kane turns that into a usable discipline.² Between two hours of writing and the claim 'this student can do legal analysis at a 68 standard' sit four distinct inferential steps:

1. Were the marks assigned consistently?
2. Would the score hold up across other questions, on other days?
3. Does performing well under these conditions tell us anything about the capability we care about?
4. Do the uses we make of the score follow from it?

The chain is only as strong as its weakest link. When I ran my own papers through it, what buckled first was the third: a two-hour handwritten sprint is a perfectly defensible test of some things, but the awkward question is whether the capability you would name as your unit's endpoint is among them. The two classic failure modes are familiar: the exam that tests only a sliver of what the unit taught; and the score that moves with things you never meant to assess (handwriting speed, panic management, fluency in the genre). Peter Knight completes the picture: assessing complex achievement is judgement, not measurement. Pretending otherwise does not make the judgement disappear; it only hides it.³

¹ Peggy L Maki, *Assessing for Learning: Building a Sustainable Commitment Across the Institution* (Stylus Publishing, 2nd ed, 2010) 163.

² Michael T Kane, 'Validating the Interpretations and Uses of Test Scores' (2013) 50(1) *Journal of Educational Measurement* 1.

³ Peter Knight, 'The Local Practices of Assessment' (2006) 31(4) *Assessment & Evaluation in Higher Education* 435.

Reliability is real, and the fix is not more rubric. Bloxham and colleagues gave experienced UK external examiners identical student work and watched them apply divergent, highly personal standards to the same scripts.⁴ You could run the experiment at your next staff meeting: same short answer, everyone marks it out of ten in silence, then compare in pairs. But then you would have to face what Bloxham's research predicts will be the result: differences of two or more marks, which compound across markers, questions and occasions into large differences in final awards.

The instinctive response to produce more detailed rubrics and more explicit criteria does not work on its own, because written criteria only acquire meaning through the personal frameworks of the people applying them.⁵ What really moves consistency is calibration: markers grading the same scripts, comparing, arguing, converging. Calibration beats codification.⁶ And before despair sets in, van der Vleuten and Schuwirth's synthesis from medical education relocates the problem: across a programme, the biggest reliability threat is not marker subjectivity but sampling — too few tasks, too few occasions.⁷ Their deeper point is the most liberating in the literature: quality is a property of the programme, not the paper. No single exam has to keep all three promises alone.

Fairness has layers, and consistency is only the thinnest. I find it useful to distinguish three layers of fairness. The first is procedural: anonymous scripts, criteria applied evenly. Danielle Tully argues that fairness-as-consistency is the thinnest available reading of the concept: necessary, but nowhere near sufficient.⁸ The second layer asks what the score actually measures. Henderson found the LSAT predicts law-school performance best precisely in time-constrained formats because speed itself, which nobody would list as a learning outcome, is doing a share of the ranking.⁹ The reading-rate evidence makes this concrete. Adults reading English non-fiction silently average 238 words a minute, when reading for ordinary comprehension.¹⁰ A student reading a fact pattern she must hold and apply is reading for recall, which the same synthesis puts nearer 100 words a minute.¹¹ Second-language readers are slower again: ten to seventeen per cent for highly proficient bilinguals, and well below that for readers who are not.¹² Thus a long, dense paper under time pressure

⁴ Susan Bloxham et al, 'External Peer Review of Assessment: An Effective Approach to Verifying Standards?' (2015) 34(6) *Higher Education Research & Development* 1069, 1074–7.

⁵ Pete Boyd and Jennifer Hill, 'Working with Rubrics' in Nicola Reimann, Ian Sadler and Jennifer Hill (eds), *Academic Standards in Higher Education* (Routledge, 2024) 169.

⁶ Ibid.

⁷ Cees P M van der Vleuten and Lambert W T Schuwirth, 'Assessing Professional Competence: From Methods to Programmes' (2005) 39(3) *Medical Education* 309.

⁸ L Danielle Tully, 'Behind the Curve: Rethinking Norm-Referenced Grading in First-Year Legal Writing Courses' (2025) 29 *Legal Writing* 1.

⁹ William D Henderson, 'The LSAT, Law School Exams, and Meritocracy: The Surprising and Undertheorized Role of Test-Taking Speed' (2004) 82 *Texas Law Review* 975. Henderson's data are American and from closed-book, heavily speeded contexts; transfer to the Australian open-book norm is a fair question. The reading-rate evidence (n 10) is not jurisdiction-specific.

¹⁰ Marc Brysbaert, 'How Many Words Do We Read per Minute? A Review and Meta-Analysis of Reading Rate' (2019) 109 *Journal of Memory and Language* 104047, 21.

¹¹ Ibid 18, 21.

¹² Ibid 19.

examines reading speed and language background alongside the law, whether or not anyone intended it. The student at the bottom of that range is not underprepared or under-able; she is simply still reading the facts while the fastest pens in the room are already writing. None of this makes timed exams indefensible; it means time pressure is a decision about what you are measuring, and that decision should be made, not inherited. The third layer asks whether the genre's tacit rules were ever taught: where exam technique goes untaught, the exam partly ranks students on their skill at guessing what each marker rewards, an equity burden that falls hardest on students furthest from the profession's cultural inside track.¹³

III Advise a novice unit coordinator on the pedagogical risks and hidden variables inherent in the inherited exam template. (30 MARKS)

The experiment is not conducted on a blank page, and would be poorer if it were. Some of the frame is institutional and blunt: at my university, a formal invigilated exam must be two hours long, with no separate reading time, handwritten. Nobody designing a paper chooses that envelope; you choose whether to enter the invigilated channel at all, and the envelope comes with the choice. Some of the frame is resources, which have a history of doing the designing: Thornton records the drift back to traditional exams under massification as logistical accommodation to marking burden, not pedagogy,¹⁴ and Rocklin finds marking capacity the most-cited reason exam skills go untaught.¹⁵ The 100 per cent final does not remove marking labour; it piles it into a post-semester spike,¹⁶ and James and Jensen's dialogue on the comprehensive final ends with four conditions of acceptability: informed students, prior formative feedback, some choice among assessment regimes, and degree-level skills safeguarded elsewhere.¹⁷ Your own unit is worth checking against all four; I did not entirely enjoy the exercise. And some of the frame is the climate: generative AI has made unsupervised assessment feel unverifiable, and the profession's regulators recently floated the idea that at least half the assessment in every Priestley 11 subject be invigilated,¹⁸ so the pull of the exam room is stronger than in decades. That pull deserves a note of its own; for

¹³ Mary Beth Beazley, 'Better Writing, Better Thinking' (2004) 10 *Legal Writing* 23; DeShun Harris, 'Let's Talk About Grading, Maybe' (2022) 45 *Seattle University Law Review* 805, 835–8; Joan M Rocklin, 'Exam-Writing Instruction in a Classroom Near You' (2018) 22 *Legal Writing* 189, 211.

¹⁴ Margaret Thornton, 'The Idea of the University and the Contemporary Legal Academy' (2004) 26 *Sydney Law Review* 481, 500–1.

¹⁵ Rocklin (n 13) 211.

¹⁶ Nick James and Darryn Jensen, 'A Dialogue concerning the Merits of the 100% Final Examination in the Assessment of Law Students' (2011) 10(3) *Canberra Law Review* 49, 50.

¹⁷ *Ibid* 79–80.

¹⁸ Legal Services Council and Law Admissions Consultative Committee, *Consultation Paper on Proposed Revisions to the Accreditation Standards for Australian Law Courses* (March 2025) section 'Invigilated assessments', proposed cl 4.7(b).

now, if regulation is about to fix how much examining we do, the design questions that remain ours matter more, not less.

So what does the inherited template actually contain? I have done my best to find out by analysing our school's available past exam papers: first the 2022–2024 papers across core and option units, coded against thirty-odd variables, then back through the archive to 2013. This gathered just over ninety papers, with question-level coding of more than two hundred questions. Three caveats: the coding is mine alone; the corpus is one school's; and I report only aggregate patterns (no unit other than my own is identifiable) because the point is self-knowledge, not scorekeeping.

The corpus is strikingly uniform, in two layers worth keeping apart. One is the institution-imposed envelope (two hours, no reading time, handwritten), which no coordinator chose and no coordinator can vary. The other layer is chosen, or at least chosen-by-default: almost no paper articulates what it is assessing, and almost none discloses criteria. Students meet paper after paper that asks for their best work without ever saying what 'best' looks like. At question level, coded against Bloom's taxonomy,¹⁹ roughly half of all questions ask students to analyse, a fifth to apply, and none to merely remember (recall having been quietly pensioned off by the open-book norm). The thinness is at the top end: questions asking students to evaluate (judge, critique, weigh a reform) make up about a quarter but live almost entirely in essay questions, a form several units have dropped; and questions asking students to create something (draft, design, propose) are close to extinct, at under one per cent. And the most arresting finding: in the core-unit problem questions, the instruction is 'advise'. It is almost always 'advise'. We tell students the degree builds critique, evaluation and reform capability; the final assessments mostly ask them to advise a client on settled doctrine, again.

IV Draft a proposal for a first-principles examination, detailing the required modifications to the assessment's small print. (40 MARKS)

Now the experiment: sit down with the brief and the constraints, refuse the filing cabinet, and derive a paper. The first surprise is convergence. First principles do not abolish the exam — some lawyer capabilities cannot be assessed by any exam format — but that makes the exam one instrument among others, not nothing.²⁰ Open book survives too, with a string attached: the closed-book orthodoxy sometimes defended as a professional requirement is a myth, and memorisation is hard to justify for a practice in which few competent lawyers work from

¹⁹ Lorin W Anderson and David R Krathwohl (eds), *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives* (Longman, 2001).

²⁰ Amanda Cahill-Ripley, 'Innovative Methods of Assessment in Law: The Value of Open-Book Exams as a Catalyst for Improving Teaching and Learning in the Law School' (2015) 49(2) *The Law Teacher* 206, 216–17.

memory alone;²¹ but open-book conditions only deliver their higher-order benefits when the teaching and the questions are redesigned for the format, not merely substituted in.²² The problem question survives scrutiny as well: it integrates issue-spotting, doctrine and application, and students prefer problem questions whatever the format.²³ However, its limits deserve equal candour: Reid classifies it as a problem with predetermined solutions and known reasoning paths, not the messy, ill-structured kind much real legal work is,²⁴ and Mertz reads the form as quietly teaching while it tests, rewarding reading for authority and flattening the human beings in the fact pattern.²⁵ So the derived paper keeps the problem question and surrounds it: perhaps with something evaluative, to reach the level the corpus undersupplies, and maybe a modest drafting task, in a genre that is unambiguously a lawyering genre. It declines the multiple-choice section, whose marking economy comes at the price of a narrowed construct and maximum ease of machine completion.²⁶

The real divergence is the small print. The biggest gap between the derived paper and the inherited one is not architectural at all: the first-principles paper *says what it is doing*. It states, on its face, what it is assessing. It discloses the criteria and the mark allocations, so that a student's split of two scarce hours is an informed decision rather than a guess about the marker's undisclosed priorities. It counts its own words against what a person can read in the time allowed, as a paper that cannot comfortably be read in the time has made speed part of the test without saying so.²⁷ And it arrives with a marking plan, not just a marking guide: a calibration step designed at the same time as the questions, because that is where consistency is actually manufactured.²⁸ Every one of these moves is cheap; none needs a rule change, a committee or a technology. Which is what makes the corpus's silence so instructive: the template's most consequential defaults are ones few would defend out loud; they persist because the filing cabinet has never contained a counter-example. The first-principles exam, in short, is recognisably our exam: invigilated, open book, two hours, anchored by a problem question. The differences that remain are unglamorous, inexpensive, and almost entirely within an individual examiner's gift. The revolution, such as it is, is in the small print.

²¹ Matthew Marin and Amanda Fisher, 'Memory: The Past, Present & Future of Law School Exams' (2023) 38(1) *Western Michigan University Cooley Law Review* 29, 29–33; Cahill-Ripley (n 20) 214–15.

²² Cahill-Ripley (n 20) 207–9.

²³ *Ibid* 215–16.

²⁴ Katie Reid, 'Using Critical Pedagogy and Legal Design Methodology to Develop the Traditional Problem Question' (2024) 58(1) *The Law Teacher* 20, 26.

²⁵ Elizabeth Mertz, *The Language of Law School: Learning to 'Think Like a Lawyer'* (Oxford University Press, 2007) 215–22.

²⁶ John Bliss, 'Teaching Law in the Age of Generative AI' (2024) 64 *Jurimetrics Journal* 111, 137.

²⁷ Brysbaert (n 10).

²⁸ Boyd and Hill (n 5).

V EXTRA CREDIT: Apply the principles of intentional assessment design to evaluate and reform your own unit's most recent final paper.

Thought experiments are cheap; semesters are not. My own unit, LAWS5102 Administrative Law, carries a two-hour, open-book final exam worth 60 per cent of the unit, and the corpus shows it oscillating between three-question and two-question architectures from 2018 to 2024 with no settled pattern — a record I now read, a little ruefully, as template inheritance in action. The first deliberate divergence came with a policy-advice question in 2024 and 2025: instead of a second multi-issue problem, one question put the student in the role of adviser to a Minister weighing discrete reform proposals to a known statutory scheme, and asked for a reasoned evaluation of each. This design sought to test structural doctrines cleanly, reach the evaluative level our papers undersupply, and reward command of a statutory framework as a system, while the retained problem question kept integrated application. The redesign has continued since, further into the small print; but this year's paper has not yet been sat, and an account of a design whose cohort has not met it would be advertising rather than evidence. I will report on it (results and regrets included) once it has run.

I invite you to join the experiment. Take your most recent paper and rate one question, one to five, against each promise: does it prompt the performance your unit outline describes? Would two colleagues mark it alike? Could a capable student who never learned the genre's unwritten rules show you what they know? Then finish one sentence: *what I most need this exam to show me about what my students know and can do is...* Now design next year's paper as its answer. If the analysis here is right, the answer will probably not require you to tear up the inherited form; it will require you to say out loud what the form has been leaving unsaid. No single paper has to keep every promise alone; what it must do is know which promises it is keeping, which it is delegating, and which constraint is binding. The exam will still end the way it always ends: a number beside each name. What design can change is how much of the truth that number carries, and how much of it you can vouch for.