
Rethinking Assessment: A Crib Sheet

Guy Claxton

guy.claxton@winchester.ac.uk

January 2021

We say we want sensitive, thoughtful, analytic, independent scholars, then treat them like Belgian geese being stuffed for *pate de foie gras*. We reward them for compliance, rather than independence; for giving the answers we have taught them rather than for challenging the conclusions we have reached; for admiring the brilliance of purely scientific advances rather than developing greater sensitivity to the inequities...we have too often ignored. George Miller¹

Here are some Points and Questions (Ps and Qs) that might help us explore better ways to evidence the benefits that young people have gained from their time in school.² They are based on a mixture of research and common sense. I think most of them are obvious, but it might be useful to have them handy as we discuss the pros and cons of various alternatives. Without cognizance of the Ps, and reasonable answers to the Qs, I don't think we can feel justified in claiming that any one method is 'better' than another. Perhaps we can collectively improve them as we go along, so please treat this as a first draft. They are in no particular order.

1. Q: What is the *purpose* of this method of evidencing capability (MOEC)?³ And *who* is it designed to help, or persuade, of *what*? What is the audience, will they find it useful, and what will they do with the evidence provided? Can the evidence be presented in such a way that others are prevented from misusing it, for example it for purposes for which it is not fit, or for making unjustified judgements and drawing illegitimate conclusions about students' capabilities?
 - P: Possible purposes include: satisfying parents' desire for information; providing useful feedback for students; meeting government/Ofsted requirements; monitoring the success of school policies/culture change; providing useful selection information

¹ George Miller, (1978). 'Teaching and learning in medical school' revisited. *Medical Education*, 12, Supplement, 120-125.

² Note how the phrasing of the purpose of education matters. We have to start by **not** using words like achievement, attainment, assessment, grades, test scores, measurement etc. all of which preempt the kinds of answers we can give to the questions that follow. See #4.

³ By using the word capability here, I am signalling my assumption that the purpose of education is to give all young people (at least some of) the knowledge, skill and character that will equip them to *do things that are likely to matter to them* (which may include reading, writing and influencing, of course); not to stuff young heads with knowledge that is inert (just because it might arguably be part of 'the best that has been thought and said').

for employers/university or college admissions. *No single MOEC fulfils all of these purposes equally well.*

- P: An increasing numbers of employers and universities are discarding school grades as not fit for their purposes. Google and Deloitte (to name but two) do not find that school grades or even class of degree are good predictors of the kinds of mental skills and attitudes they are often looking for.⁴
 - P: When the purpose is to guide and encourage pupils' development, rather than to accredit attainment, then clearly formative kinds of 'assessment' are to be preferred to summative. As Dylan Wiliam points out, aircraft fly by constantly monitoring their direction and making small adjustments; they don't wait till they have flown 3000 miles before checking whether they are on course.
 - Ipsative assessment – gauging progress in terms of personal progress and improvement, rather than comparing performance against fixed benchmarks (criterion-referenced testing) or against a population of peers (norm-referenced testing) – is much more effective at promoting pupil engagement and improvement than any kind of summative testing.⁵
2. Q: Have we considered the full *range* of potential MOECs before settling on 'the best one' (for a particular purpose)? It might be a written test; a questionnaire; a teacher appraisal; a representative e-portfolio that points to accomplishments and dispositions; a written self-reflection; a 360° appraisal orchestrated by a student; a viva; a performance of competence; performance in a designed problem situation; and so on. Or some combination of these.
- P: All of these can be done well or badly, or appropriately or inappropriately. *Quantitative measures are not necessarily better, more objective or more revealing than qualitative.* Numbers can conceal subjectivity, and can be applied well or badly, appropriately or inappropriately.
3. Q: Do we have a particular *prototype* at the back of our minds when thinking about MOECs? Would it be good for *this* situation to design a test that is a bit like: getting a badge in the scouts; a PhD viva; a driving test; an interview for a promotion; the IB Diploma; the (soon to be defunct) Cambridge Pre-U; a poetry or film competition judged by experts; a critical review of a book or play; a Grade 3 piano exam; an IQ test; submitting an article to a refereed journal; a competition like *Masterchef* or *University Challenge*?
- P: Maths and Physics are not necessarily good prototypes for thinking about evidencing in other subjects – just because they are more likely to have 'right answers' that are (a) unequivocal, and (b) countable. Of course, you can make English, Drama or Design Technology more like Maths by shrinking them to focus on prescribed definitions, lists of spellings, and 'formulae' (like grammar), but if you do so, much that is characteristic and valuable about those disciplines is eradicated.
 - P: (Even Maths and Science involve much more than Right Answers and Correct Calculations. School maths and science have become 'cartoon' versions of 'real' maths and science, showing little of the essential struggles with technology, fierce

⁴ David Goodhart's recent book *Head, Hand, Heart* has good discussions of this.

⁵ See the classic paper by Terry Crooks, *The Impact of Classroom Evaluation Practices on Students*, 1988, *Review of Educational Research*, 58(4):438-481.

professional rivalries and disputes, ingenuity and imagination in theory-building, temptations to cheat; or frustration and exhilaration.)

4. Q: Have we got the right *word* to describe the *kind* of test that we think is appropriate? Are we *assessing, evaluating, measuring, testing, illustrating, tracking, demonstrating, or evidencing*? Each of these words suggests a particular kind of assessment – which may or may not be appropriate.
 - P: It is important not to allow ourselves to be inadvertently trapped into choosing or valuing certain kinds of MOEC prematurely, simply on the basis of the words we are using. (*I prefer evidencing as the most open-minded and non-prejudicial.*)
5. Q: When and where would be the right *time* to apply an MOEC? Would the MOEC be better situated at an *entry point* – e.g. to a college or university course or to employment – rather than at the *exit point* from school?
 - Q: How do we weight up the pros (of which many are self-evident) and cons of such a suggestion?
 - P: In many cases it seems quite easy to ascertain someone's level of literacy, numeracy, critical thinking, ability to think on their feet etc. with short tests at the entry point, and such tests can readily be tailored to the level and nature of the expertise that the new 'job' demands. And it would be harder to cheat.
 - Q: But then on what basis do we assess the effectiveness of schools?
6. Q: Should the timing of a test be decided by *age* or by *readiness*? And if by readiness, is that judgement made by the learner, by a teacher, or by both?
 - P: Your driving instructor usually has a major say in whether you are ready to 'go in for the test'. In the scouts or guides, you may decide for yourself when you are ready to go for a badge.
7. Q: What range of *ages, abilities, aptitudes and attitudes* is the MOEC appropriate for? What is the evidence? Does a single test enable valid and useful discriminations to be made across the full range and diversity of the intended student populations? Is it equally 'fair' to all students?
 - P: This problem is radically reduced if you adopt the Entry rather than Exit strategy.
8. Q: How *trustworthy* is the MOEC? Is it genuinely 'objective' and proof against contamination by e.g. cheating or the (conscious or unconscious) bias of an examiner?
 - P: We know that all kinds of judgements – both in marking of scripts and 'teacher assessments' - can be subject to such biases. Studies show that even irrelevant student features such as names, photos, and handwriting can contaminate such judgements.
 - Q: Is there a trade-off between the (apparent) objectivity and the predictive validity of the test? (In other words, does increasing objectivity tend towards making tests more sterile and artificial?) Is there a point at which the 'cure' becomes worse than the 'disease'?
9. Q: Does performance on the MOEC depend upon *extrinsic factors*: considerations other than those which you want to be evidencing? If so, it is not a good index of the target capabilities.

- P: For example, many tests are carried out under high levels of emotional and time pressure (the stakes are high, and you have to work fast to do well) – so they do not provide a valid index of how well you can access and apply your knowledge under less stressful conditions.
 - P: Performance under such conditions relies to a significant (but unknowable, and therefore not discountable) extent on ‘exam-craft’ (e.g. reading the examiner’s mind; apportioning time; mastering the nuances of expression that gain marks), and on the absence of psychological worries and distractors and/or the ability to manage those that are present.⁶
 - P: Studies have shown that students’ test performance depends to a very significant extent not on what students actually know (or can do) but on whether they perceive that this problem actually calls for that piece of knowledge. They often know it, but it doesn’t come to mind when it should. (Q: What are the implications of this for (a) testing, and (b) teaching?)⁷
10. Q: What effects do particular MOECs have on the school *staff*, especially teachers and school leaders, on their pedagogical style, and on the working of *exam boards*, *inspection regimes* and so on? The nature (and status) of tests drive the way that teachers teach, and inevitably lures school leaders towards ‘gaming the system’ in the interests of their school’s reputation, or even salvaging their own careers. Syllabuses get designed around what is mark-able (rather than what is remarkable, i.e. interesting and challenging) in particular subjects.
- P: These distorting effects of different MOECs are not evidence of individual weakness or professional laxity; they are inevitable consequences of the implicit incentives and sanctions that are built in to a model of assessment.
 - P: All of these unintended costs and consequences need to be factored in, in deciding whether, overall, a particular MOEC is accurate, beneficial, and fit for purpose.
11. Q: And what effects does the MOEC have on the *students*? How does the anticipation of a particular kind of test influence students’ motivation, and their methods and styles of cognitive engagement? E.g. do they judge that, to do well on the test, it is better to opt for near-verbatim retention, and the ability to perform accurate computations, or to aim for deeper and more challenging forms of understanding?
- P: There is evidence that boys are more willing to adopt this expedient kind of learning style than girls, who may consequently be disadvantaged by their intellectual integrity and curiosity in those kinds of tests.
 - P: There are many ways in which ‘the tail of assessment wags the dog of learning’: the nature of the anticipated MOEC powerfully influences students’ engagement, attention, cognitive and learning styles – for good or ill.

⁶ I have a text sent to me by a young woman, now at Cambridge, setting out in excruciating detail exactly what it takes to get an A in History A-level, and the deleterious effects that these techniques had on her own inquisitive learning.

⁷ See, for example, David Perkins, *Outsmarting IQ*.

12. Q: Do MOECs that are designed to evidence one set of desirable outcomes of education (e.g. disciplined knowledge) have any unintentional impact on other desirable outcomes (e.g. the development of resilience or initiative) that are not being captured by that MOEC?

- P: For many people (including me) the desired outcomes of education (DOEs) fall into three broad categories: what we want young people to *know* (knowledge and understanding); be able to *do* (literacies and disciplinary expertise); and *be like* (character, attitudes and dispositions). I see these as *layers of learning that are going on in every classroom, all the time* (not, as some do, as forms of content that compete for time and attention, so that developing ‘skills’ or ‘character’ must necessarily result in the neglect of Shakespeare and calculus).⁸ On this ‘infused’ view, different pedagogies, pressures, and classroom cultures may all be effective at developing knowledge, but have quite different effects on the development particularly of *epistemic character* – the set of attitudes that underpin a person’s confidence and capability in dealing things that are complex, demanding, challenging or uncertain, within the school curriculum but, much more importantly, in out-of-school life.⁹
- P: Studies have shown, for example, that Direct Instruction with constant testing and repetition, in which everything is explained and practised, and nothing is left for students to explore or discover, is effective at developing certain kinds of knowledge and disciplinary skill, but at the cost of weakening curiosity and creativity. Knowledge-based testing can encourage attitudes of conformity and correctness rather than exploratory or imaginative thinking. It is a value judgement whether this collateral damage is a price worth paying for good grades.¹⁰

13. Q: What are the desirable habits of mind that we might want all young people to develop at school, and how is their growth best evidenced?

- P: James Heckman, Angela Duckworth and others have shown that success in life, as judged by a whole array of socioeconomic indicators, depends more on the possession of certain character traits than on academic qualifications. These include perseverance, curiosity, open-mindedness, intellectual humility, rational scepticism, collaboration, and empathy. There is much current work on devising ways to evidence these that do not impact negatively on more traditional (and easier to measure) outcomes.¹¹
- P: Through neglect of this implicit character-shaping, schools can turn out young people, with good or bad results, who are, for example,
 - Timid, compliant, and dependent, or
 - Glib, smug, arrogant and intellectually pugnacious, or
 - Defeated, resentful, anti-social and anti-intellectual¹²
- Q: “If we don’t find ways to measure what we value, we simply end up valuing what we can measure.” If things that are not ‘assessed’ have less status than those that

⁸ See Guy Claxton, *The Future of Teaching and the Myths that Hold it Back*, Routledge, 2021.

⁹ See Guy Claxton, *The Learning Power Approach: Teaching Learners to Teach Themselves*, Corwin, 2018.

¹⁰ See especially Yong Zhao, *What Works May Hurt: Side Effects of Education*, Teachers College Press, 2018.

¹¹ See, for example, National Academies of Sciences, Engineering, and Medicine, 2017. *Approaches to the Development of Character: Proceedings of a Workshop*. Washington, DC: The National Academies Press.

<https://doi.org/10.17226/24684>. Especially Chapter 6, *Measuring Character*, pp63-75.

¹² You will readily be able to provide your own examples of these types as they progress into adulthood.

are, how important is it to incorporate ways of evidencing epistemic character development into whatever assessment regime we recommend?

- P: Despite the use of the word 'measure' in the previous bullet point, it is imperative that the 'hegemony of the quantitative' be resisted in thinking about how to evidence the development of students' habits of mind. Qualitative judgements are perfectly well accepted in the big wide world, especially in the hard-nosed business world, where annual reviews and 360 degree appraisals are *de rigueur*, so why on earth should they be treated with such suspicion in the world of education?

14. Q: Is testing knowledge and comprehension a good way of predicting competence? As I said at the beginning, it seems trivial to point out that the value of education lies in the development of competence: the ability to get things done that matter. The obvious way to see if someone possesses the ability to do something is to ask them to do it – not to talk or write about it. And especially not if comprehension turns out not to be a reliable indicator of competence.

- P: This is the elephant in the examination hall. Much testing in schools depends on the assumption that comprehension is prior to, and necessary for, competence. You have to understand something before you can do it; and showing that you understand something is sufficient to reassure someone that you are able to do it. But this tight association between comprehension and competence is patently false. I can tell you a lot about rugby (for example), but would be worse than useless on the field. A good theatre critic need not be able to act or write plays. So displays of understanding are no guarantee that you can practise what you are preaching.

On the other hand, much expertise is unconscious and incapable of articulation. Indeed, like the ability to walk or dance, you may *never* have been able to describe or explain what you are doing. You just picked it up. And this lack of ability to explicate your expertise is no handicap at all in many real-life fields of activity. Only when you are put into the role of teacher, coach, mentor – or in a decision-making meeting – does the ability to unpack and articulate your competence become a necessary part of the skill. You use words to draw your learners' attention to valuable areas (or minutiae) of their experience, and then practice and awareness do the rest. Or you join a discussion in order to pool the pros and cons of various courses of action, or to organise your various roles in a joint endeavour.

Jeanne Bamberger ran a 'Lab for Making Things' at Harvard in the 1980s in which children came to solve practical problems in pursuit of making things that worked. Like a gate that opened and shut or a mobile that rotated and balanced. She found there were many children who sat at the opposite ends of a spectrum. There were those who could be successful at these practical tasks, but could not explain the principles behind their success; and those who were good at explaining the ideas, but who, for the life of them, could not make their mobiles balance. Comprehension without competence, and competence without comprehension, were both

common. The latter children would have been grievously, and unjustly, penalised if the only test available was a written one.¹³

- P: As Ron Berger, Chief Education Officer of the EL Education schools in the USA, is fond of saying: for the bulk of our lives, most of the time, we will be judged not on our ability to knock out small essays, or to perform correct calculations (that a machine will do faster and more reliably), but on the quality of our work (real work, not school-work) and of our character: on our behaviour, not our protestations. Any acceptable method of evidencing capability must start from and incorporate this reality.
 - Conclusion: comprehension (and displays thereof) are neither necessary nor sufficient for the development of capability, and are often deeply misleading.
15. P: Politicians are mostly stupefied (“rendered unable to think clearly, especially about complex or delicate matters”) by the nature of the culture they inhabit: adversarial, tribal, hyper-sensitive to popular opinion, beholden to wealthy and opinionated lobbyists, and fixated on creating short-term appearances of success. This mindset disables their ability to think deeply and productively about the problem of educational assessment (with a few honourable exceptions). They will not see that there are strong reasons why traditional exams, even if tinkered with, are massively unfit for purpose. It is therefore down to the profession – those in touch with the complex realities and entangled moralities of children’s schooling – to take a lead on pressing for radical change in the ways that the benefits of education can best be evidenced.

¹³ Jeanne Bamberger, The laboratory for making things: Developing multiple representations of knowledge. In D. A. Schön (Ed.), 1991, *The reflective turn: Case studies in and on educational practice* (pp. 37–62). New York, NY: Teachers College Press