A critical comparison of eleven assessment types

(From 'Making Learning Happen: 3rd edition', Phil Race (2014) London: Sage)

The table below illustrates the pros and cons of several kinds of assessment. In each case, in the 'status' column I have included judgements regarding how well (or how poorly) each assessment type listed measures up to validity, fairness, whodunit? (whether there could be serious doubts or not regarding whose work is being assessed), links to the real world, and the extent to which feedback to learners may be available, or useful.

Type of assessment	Status	Advantages	Disadvantages
1 Traditional Exams Exams are often referred to as the 'gold standard' because of their widespread use in secondary and higher education. Still the most common kinds of exam, are handwritten, invigilated, against-the-clock, with questions not being known by candidates beforehand. Exams remain prevalent on many post-compulsory education courses, sometimes where questions are set by external examiners, and sometimes by the staff who teach the learners. Timescales vary, but two hours and three hours are relatively common in universities, though much shorter exams are perfectly possible (and perhaps desirable). Many traditional exams offer candidates a choice of questions (e.g. attempt any 5 out of 8 questions, each carrying equal marks), but increasingly there may be a compulsory section, then a section providing choices.	Validity: poor, limited to what comes out of pens. Fairness: can be good, but poor when answers are essay-type, and different markers would award very different marks for the same essay. Whodunit?: relatively safe (though stories of ingenious cheating are legion!). Real world: written exams are not at all close to the workings of the real world; most people never do a written exam again after leaving university. Feedback to learners: very limited indeed, usually just a score/grade or pass/fail, which can leave candidates having very little idea about what they did well or badly.	Can avoid plagiarism and cheating. Give data which can be ranked and handled quantitatively. Exams are relatively familiar to learners entering higher education, as they've already experienced them at school. Exams are already 'hard-wired' into many university systems, so a case doesn't have to be made for continuing to use them. Written exams are much better for some subjects than others: for example they can work well for mathematical and quantitative subject matter, and tend to work really badly for 'wordy' or descriptive matter.	 As has been argued already in this book, exams tend only to measure what comes out of pens, a poor proxy for what might be in heads. Many otherwise capable learners never show their best efforts under exam conditions. It can take a long time to mark a set of exam scripts (properly). (There are economies of scale for large numbers of candidates as examiners become familiar with the marking scheme being used). Problems with speed of writing and legibility, and difficulties candidates face when using a se. 'Sudden death': a bad day can mar a lifetime. A snapshot of achievement, rather than a real measure of it. One of the main skills measured tends to be time-management - dividing the available time sensibly between the questions being attempted. Promotes surface-learning: filling heads with information to use 'on the day' and forget as quickly as possible thereafter. Question-spotting by candidates can pay off substantially, meaning that at least some candidates pass without having learned the whole syllabus reasonably well. Where candidates have a choice of questions, it is really hard to get all of the questions to be of equal difficulty - leading to at least some candidates ending up with an easier exam than others overall. Long-answer written exams contribute to the continuing trend for post-compulsory education to remain 'elitist' - i.e. to favour those who are good at such exams.

2 Short-answer exams Short written responses to a large number of questions. Usually the whole exam is compulsory, reducing the tendency for candidates to us question- spotting as a means of deliberately only learning some of the syllabus.	Validity: often a lot better than long-answer exams, as evidence of achievement of learning outcomes can be covered much more fully, rather than the 'write down everything you happen to know' tendency which long-answer questions can engender. Fairness: can be good. Whodunit?: safe under exam conditions. Real world: closer than traditional exams, as success may involve knowing the overall subject really well. Feedback to learners: poor, usually just a score or grade.	 Can cover a wide range of topics in a limited time. Not so much affected by speed of writing or legibility. Can be somewhat faster to mark than long-answer exams (but not always). Can measure breadth of knowledge. Is fairer than exams where there are choices of question, as all candidates are effectively taking exactly the same exam. 	 Can miss out on depth of knowledge. Can deprive high-fliers of the opportunity to excel. It can take much longer to design a short-answer exam paper than a traditional long-answer one. It can be quite difficult to apportion marks across the various elements in a short-answer exam paper.
 3 Multiple-choice exams Paper-based or computer-based: e.g. select the 'best' option from 4 or 5 alternatives for each of a fairly long list of questions. Such exams are usually time-constrained, but often most candidates complete the exam with plenty of time to spare. Note that these can be extended to be multiple-response exams, where each question is along the lines of 'Which (one or more) of the following options is true?'. This can cause candidates to think harder about <i>all</i> of the options, rather than just picking the 'best' then moving on. 	 Validity: can be good if questions well-designed. A wide range of syllabus knowledge can be addressed in quite a short time. Fairness: can be really good, if questions well designed and trialled. Whodunit?: safe under exam conditions; not safe if asynchronous, for example in distance learning contexts. Real world: can be better than other kinds of exams, as multiple-choice exams tend to measure what goes on in heads, and aren't limited by what comes out of pens – they can measure decision making well. Feedback to learners: possible to be excellent in speed and quality (but not often achieved). Excellent feedback can address, for each option: 'was I right?' and (particularly) 'if not, why not?' 	 Gets away from 'what comes out of pens' limitations of other kinds of exams. When questions well-designed, can quickly test quite a wide range of subject knowledge. Can be useful in areas where rapid decision- making is a useful skill for learners. Possible to provide feedback on-screen after each decision in computer-based uses, or a complete feedback printout (feedback on the distractors as well as the correct choices) on leaving the exam. (Such printouts can make really useful revision tools for future candidates) Can be useful for candidates who have difficulty stringing together fluent prose in written answers, but who can still think clearly through options. 	It is much harder than people think to design really good multiple-choice questions. The 'key' is the best option, and is the one intended to be chosen by candidates who know the topic properly. However, It is sometimes hard to design a 'key' which is <i>always</i> right – high-fliers can often spot when even this choice is not correct. It can be difficult to design good 'distractors'. Options such as 'all of the above' or 'none of the above' are still too often included (usually though laziness regarding thinking of more distractors), and which are very rarely the 'best' option in any case. Questions need to be well piloted and tested before being used in exams. Still an element of 'luck' picking the best option (in a test using four-option questions, the average monkey should score 25%). Quite a lot of emphasis now goes onto <i>reading</i> the questions and options well; learners with limited skills or speed in reading can be disadvantaged.

4 Essays	Validity: rarely good (prose which comes out of a pen in an exam,	Allow learners to	Essays take forever to mark, and marking is unreliable
Essays are highly regarded as an assessment type, despite the many disadvantages listed in this table! At university level, a coursework essay is often an extended account of length such as 3000 words, usually submitted in word-processed form, and often passed through plagiarism- detection software. However, handwritten essays (shorter) in exams are still widely used. In the case of coursework essays, learners may be given an essay title or theme, or may be allowed to choose from a list of topics, or may have leeway to choose an appropriate topic of their own, or negotiate a topic with a tutor.	or through a keyboard in coursework, is rarely the best way of measuring evidence of achievement of intended outcomes). Fairness: poor (a great deal of research evidence is available showing that different markers award very different marks). Whodunit?: very unsafe, except for essays under exam conditions. Essays can be commissioned and purchased online from well-practised skilled writers! However, concerns about Whodunit? are minimal with small groups of learners, where tutors regularly talk to learners and would usually quickly know if submitted work was not their own. Real world: not close to the sorts of writing relevant to most careers. Feedback to learners: can be useful, but usually too late, not least because of the length of time it takes to mark a set of essays. There is usually no feedback at all on essays handwritten in exams.	demonstrate ability to construct written arguments, and to write fluently. Can give candidates who show 'depth' fair reward. Can give an indication of the quality, depth and breadth of reading that has been done by candidates.	 (unfair) anyway, as proved by a great deal of research! Tends to advantage learners who are good at written 'waffling'! Unless there are tight word-limits, a longer essay will usually score higher than a shorter one. Where there is an element of choice (e.g. coursework essays) some choices may prove harder to bring off in practice than others, disadvantaging some learners. In coursework essays, there can be a tendency to copy in (suitably rephrased) sentences from literature sources, without really thinking about the meaning of the elements copied in. Spelling, punctuation and grammar may disproportionately affect marking. 'Coherence', flow, ease of reading essays disproportionately influences most markers. A 'smooth' essay is usually awarded higher marks than a 'jerky' one, even if the content of the latter is much better. Where essays are handwritten in exams, it is not at all easy to edit and adjust along the way, e.g. to go back and rephrase the start of the essay appropriately after the main thrust has been addressed. Handwritten essays in exams are subject to concerns about measuring 'what comes out of a pen' rather than 'what's in a head', and are subject to the effects of speed of writing, legibility and so on. Handwriting an essay in an exam is quite a different game than composing a word-processed coursework essay, so coursework is poor preparation for the exam experience, and feedback on coursework essay may not help exam candidates.

 5 Annotated bibliographies This kind of assessment works best when various elements of the task are specified, including one or more of: Overall word count, preferably with quite tight limits; Number of sources to include – preferably an exact number in practice; Any expected balance between kinds of sources, e.g. journal articles, reviews, book chapters, web sources. Whether to include (say) four 'given' sources, and four more that individual learners have found; Whether to prioritise the sources in order of any particular aspect: e.g. usefulness, authority, relevance to topic, and so on; Whether the list should be a comparative one, e.g. with pros and cons of each source in the context of the bibliography; Whether there is free choice regarding the 'age' of the sources. For example, in some contexts it can be useful to ask limit contents to sources published within the last 3 years. The extent to which the comments on each item should include the learner's own view. 	 Validity: can link well to learning outcomes about breadth of reading, and prioritising quality of sources. Fairness: can be much better than essays or reports, as high-flying candidates quickly distinguish themselves by the quality of their comments on sources, and their rationale for the sources they choose. Whodunit?: good, as any unwanted collaboration or copying would be fairly obvious (e.g. same sources selected, in same order, and with identical mistakes in the referencing!). Real world: high relevance to many careers, where learners will need to be able to review a range of sources and select and justify those most relevant to a given context. Feedback to learners: this can be harder to achieve than with some other assessment types, but can be really valuable when done well (e.g. it would have been good for you to have included source 'x' because, source 'y' was not a good one to include, because, and so on). 	 A really useful way to cause learners to read around a topic, rather than just dip into random sources during their studies. 'Wikipedia' (for example) could be allowed to be linked to only one of several required sources. Learners can be given the opportunity to demonstrate the breadth of their learning (range of sources) as well as depth (their judgemental comments about respective sources). It can be useful if one of the assessment criteria relates directly to the correct 'citing' of each source (e.g. 'Harvard system'), so that learners gain practice in getting references exactly right, which may be important for future research-related writing. A set of annotated bibliographies can be retained as an online resource, to show future cohorts of learners how to do this task, and for these learners to practise making judgements to establish criteria for their own work. A '300-word annotated bibliography prioritising five sources' can be far faster to mark than would be an essay where similar literature reviewing was intended – and can be marked much more reliably (fairly) than such an essay. 	The easiest sources to find tend to be via Google (especially the first page in a search) and Wikipedia, and it may be necessary to prevent learners from making more than limited use of these sources. Annotated bibliographies essentially require learners to demonstrate their academic literacies, but any lack of relevant information literacies may get in the way of this. There is the danger that the extent of sources chosen by learners may mask the depth of thinking about individual sources, therefore it is important not just to get learners listing a lot of sources. Learners may need some rehearsal before undertaking this sort of task for assessment. For example, a whole-class session could be taken up with learners assessing some past examples of annotated bibliographies, so that they found out how the assessment worked in practice before making their own contributions to the genre.
6 Reports For example, write-ups of practical work, field work, investigations, and so on. Usually word-processed these days.	 Validity: can be reasonably high. Fairness: not great, but much better than essays. Can allow room for good candidates to shine. Whodunit?: can be unsafe, unless other ways of checking, e.g. face-to-face quizzing. Real world: can be good: report writing relevant to many careers. Feedback to learners: can be useful, but usually comes too late. 	 Avoids 'sudden death' aspects of assessment, as reports are usually built up over a period of time. Assessment can be broken down usefully, such as agreed proportions of marks for 'Abstract', 'Method', 'Interpretation of Data', 'Conclusions', 'Plans for further work' and so on. Can be on work done collaboratively, but with individual write-up. Learners who write-up quickly (before they have forgotten what they actually did) can be advantaged, therefore encouraging good study habits of 'keeping up' and 'avoiding backlogs'. 	 Word-limit may need to be controlled strictly, as long-reports would otherwise almost always score more marks than short ones, whereas in the real world a really good short report may be much more useful in practice. There is the danger that learners can spend more time on writing reports, than is reflected by the marks they carry overall in the bigger picture of assessment. Those learners who end up with a backlog of reports may spend far too much time catching up on this backlog at the expense of preparing for summative exams, which may carry much more weight in the overall assessment.

7 Portfolios of evidence For example, built up over a period of time on a course or module, often with intermediate feedback opportunities for learners. In the present climate of 'evidence-based practice', assessment by portfolios is naturally gaining momentum. As can be seen in this table, the 'status' aspects (validty, etc.) are favourable, and there are many advantages, but the crunch comes in the disadvantages column – portfolios take a great deal of time to mark, and there are distinct problems regarding fairness of assessment.	Validity: can be good, as different elements of a portfolio can relate to each different aspect of evidence of achievement of a range of learning outcomes. Fairness: can be good, but different assessors may be looking for different things in a portfolio, in which case fairness can be poorer. Whodunit?: can be questionable, but improved where face-to-face probing is also used. Real world: can be better than many other kinds of assessment, depending on what's included in the portfolio specification. Feedback to learners: can be good, especially if their progress is reviewed at various times during construction of a portfolio.	 Allows for a wide range of kinds of evidence of achievement, for example drawings, photos, videos, recordings, reviews, reflective commentaries. Can extend well beyond the 'read-write' domain. Can build in opportunities for learners to reflect on their learning, and provide evidence of such reflection. Can be interdisciplinary, helping learners to link together aspects of different subjects and topics. Allows candidates to demonstrate originality and creativity. Portfolios can be useful evidence to show prospective employers. Portfolios can be maintained and updated beyond the assessment period. 	 It can take forever to mark a set of portfolios, and they can be very bulky to carry around from one marking place to another (work to home, and so on). Can be difficult to balance marks for portfolios evidencing different strengths. A big portfolio will normally attract higher marks than a small one, so there is a tendency to reward 'cramming in as much as possible' rather than quality of evidence. One of the most significant dangers with such a 'big' assessment element is non-completion. When portfolio assessment is used, the portfolio may often make up a substantial part of the overall assessment, which can mean for those candidates who are not at their best in this element, assessment is unduly prejudiced against them.
8 Oral exams These are essentially about interrogating individual learners face-to- face, often with two or three assessors present. Sometimes the questions may be about the course or module in general, and/or about a particular submitted element of coursework (for example a portfolio or dissertation). An oral exam does not necessarily have to be lengthy. Even a 5- minute oral exam can be useful to gain assurance about the 'whodunit?' aspect of a portfolio or dissertation. (Note that in many parts of the world there is much more oral assessment than written assessment, as was the case in the UK before written exams started in 1791 towards becoming endemic!	 Validity can be high, but still can favour candidates who can 'talk well' over those who know it just as well but are not so good at 'talking it'. Fairness: can be good when successive candidates are asked exactly the same questions (but 'security' of questions needs to be safeguarded, so they don't 'leak' to later candidates). Whodunit?: one of the safest kinds of assessment. Real world: strong links to the sorts of questioning learners will need to be able to handle in most careers. Feedback to learners: can be quick and useful, but also can be stressful and rather transient (e.g. quickly forgotten). 	 Allows for probing questions, to test real mastery rather than surface learning. Gets over reservations about 'Whodunit?'! Learning payoff during preparation for an oral exam can be wider, deeper and better than (for example) just writing an essay or report, as anything may be asked. When learners are encouraged to engage in practice and rehearsal, they can learn a great deal from each other, as well as improving their skills relating to oral performance. 	 Some candidates can be let down by nerves. With a large class of learners, a round of 'orals' can take a lot of time, and 'the word inevitably gets around' about the sort of experience it is going to be. Evidence of achievement as demonstrated orally is somewhat ephemeral, and it can be difficult to think back over several candidates performance and remember accurately who did better or worse. It can be difficult to guarantee fairness, when different degrees of probing may have been used with different candidates.

 9 Individual presentations Typically, assessed presentations are given before a relatively small group rather than a whole large class, in the presence of a tutor (who assesses) but often some peer-assessment is built in as well. Sometimes presentations are made by groups, but then it is much harder to allocate credit appropriately to individuals, so for assessed presentations, individual performance is preferable. It is useful if the briefing is really clear to learners in advance, including: The duration A range of topics to choose from - or the chance to choose something original Any guidance about supporting materials (e.g. slides, handouts, exhibits). The assessment criteria The extent of any 'question and answer' episode after each presentation. 	 Validity: can be good when learning outcomes include oral communication skills. Fairness: it can be difficult to maintain fairness during a set of presentations, as there is a tendency for later candidates to learn from earlier ones, and do better. Whodunit?: one of the safest forms of assessment. Real world: high relevance; in many careers candidates will need the skills involved in giving presentations. Feedback to learners: possible to give quick and useful feedback, but his may be somewhat ephemeral and quickly forgotten. Peer-feedback during rehearsal can however be really valuable. 	 Allows oral communication skills to be demonstrated alongside mastery of subject matter. Can allow candidates as much time as they need for preparation and rehearsal. Depth of learning tends to be high; learners tend to remember very well things they researched and practised as preparation for a presentation. Can be used in a peer-assessment context, there learners can gain a lot from making informed judgements on each others' presentations. Can allow individual learners to demonstrate particular strengths. Can include the opportunity for 'probing' to test depth of knowledge, where questions are posed for a few minutes after each presentation. 	 It can take forever to assess a large number of presentations. Choice of topic can affect marks significantly. What seemed like an interesting and stimulating topic can end up being harder than imagined, and so on. There can be some drift in standards, where 'later' candidates are judged more rigorously than 'earlier' ones, or conversely benefit themselves from seeing earlier presentations. Some candidates can be unduly disadvantaged by nerves. Impression marks associated with the quality of slides or handout materials used during the presentations may overshadow the quality of the actual mastery of the topic concerned. It is sometimes difficult to collect evidence to put forward for moderation (for example external examiner scrutiny), though recordings can be made for this purpose.
10 Posters For example, preparation of a visual display in a specified format, on (e.g.) an A1 sheet, using photos, drawings and text to address a particular brief.	 Validity: can often be high, allowing good links between evidence of achievement and intended learning outcomes. Fairness: likely to be at least some subjectivity when it comes to judgements, but this can be offset by having multiple judgements (possibly peer-assessment, and assessment by externals). Whodunit?: questionable, as learners may use varying amounts of external help in producing posters, and may collaborate with each other. The 'Whodunit?' aspect can be made much better when the assessment also included learners 'talking about their poster' or 'being questioned about their poster'. Real world: visual displays are used in many professions, for example to back-up a proposal, or present findings to colleagues. Feedback to learners: can be very effective, especially when tutor and peer-feedback is offered on drafts in class (i.e. nothing 'secret') before the preparation of the final submission. 	Can allow learners flexibility in choice, where they have significant control of the topic and the way they present their findings. Gives room for learners to organise their thinking visually rather than in words alone. Exhibitions of posters can be kept online, and used for future learners as indications of the kinds of evidence they may aim to emulate, and (better) for getting future learners to learn by assessing, before they set out to make their own posters.	It can be really hard to make relative assessment judgements about different topics handled in different ways to different depths. Wealth may come into the picture; learners who can afford good colour printing and photos may be advantaged over those whose resources are more limited. The visual aspects of the poster can dominate too much when being assessed. Judgements on visual evidence such as in posters is always to some extent subjective, with different assessors looking for different things in a 'good' poster. (This problem can be offset by having several assessors, or including one or more externals in the assessment).

11 Artefacts These can include paintings, designs, models, sculptures, items of metalwork, engineering outputs, teaching materials, plans, accounts, prototypes, furniture, display items and so on.	 Validity: this can be high, where the intended outcome of a particular curriculum element includes the production of specific items. Fairness: this can be harder to achieve in assessment of artefacts, as originality and creativity are likely to be among the assessment criteria, and therefore there is bound to be at least some subjectivity in the assessment. Whodunit? On this dimension, artefacts are relatively safe, although there remain possibilities for others to have helped in their production. Real world: artefacts often link strongly to the sorts of skills which learners need for particular vocations or careers. Feedback to learners: this can be done well, for example by 'things I like best about this example' and 'one suggestion which could have improved this example would have been'. 	 When a course or module involves learners in practical work in workshops or studios, it makes a big difference if the quality of their work there counts towards their overall assessment. The competitiveness which is encouraged by some sort of measure of the quality of the things learners make encourages them to put more effort into their practical work. Artefacts can be retained by learners after assessment, and can be useful evidence of their achievement to show prospective employers. A photographic record of assessed artefacts (or the artefacts themselves if not needed by their creators) can provide the next cohort of learners with valuable targets to aim towards – and exceed. 	 Where some learners may have benefited from external help in the production of artefacts, the fairness of assessment can be compromised. When individuals are to be assessed on the basis of artefacts they produce, collaboration between learners is discouraged, and may deprive them of things they could have learned from and with each other. It can sometimes be difficult to work out how much the assessment of artefacts should contribute to the overall assessment of a particular curriculum element. Where some learners have special needs which limit how well they can produce particular kinds of artefact, it can be difficult to make 'reasonable adjustments' to allow them alternative assessment possibilities.
--	---	--	--