A contribution to the history of assessment: how a conversation simulator redeems Socratic method

Robert Nelson & Phillip Dawson

Office of the Pro Vice-Chancellor (Learning and Teaching), Monash University, Melbourne, Australia

Assessment in education is a recent phenomenon. Although there were counterparts in former epochs, the term assessment only began to be spoken about in education after the second world war; and since that time, the views, strategies and concerns over assessment have proliferated according to an uncomfortable dynamic. We fear that increasingly, education is assessment-led rather than learning-led and 'counter to what is desired' in an ugly judgemental spirit whose moral underpinnings deserve scrutiny. In this article, we seek to historicize assessment and the anxieties of credentializing students. Through this longer history, we present a philosophy of assessment which underlies the development of a new method in assessment-as-learning. We hope that our development of a conversation simulator helps restore the innocence of education as learning-led while still delivering on the incumbencies of assessment.

Keywords: history of assessment; philology; Socratic method; conversation simulation

The language of assessment

As a term, assessment is difficult to define (Joughin, 2009); its meanings, purposes and motivations can be even more difficult to discover (Boud, 2000; Dawson et al., 2013). For clarity and inclusivity we have used Joughin's minimalist definition of assessment in our philological investigation:

> To assess is to make judgements about students' work, inferring from this what they have the capacity to do in the assessed domain, and thus what they know, value, or are capable of doing

> (Joughin, 2009, p. 16)

Joughin's definition is broad and enlightened; it represents recent thinking about what assessment *is* and what it *could be*. In this article we explore what assessment *was* through the history of assessment, with a focus on language. We find etymologies more aligned with taxation and torture than modern assessment practices, but also find the term assessment to offer opportunities as a relatively new word in this context. We conclude by proposing a scalable electronic return to Socratic dialogue.

The newness of assessment as a term in education is not just a linguistic accident. In many languages, we struggle to find a word for assessment. If we look in ancient Greek, for example— that language that gives us words like school, academy, gymnasium, pedagogy, method, syllabus and system—we find no obvious equivalent to assessment. A Greek might speak of an examination (εξέτασις) but this was applied to legal (Demosthenes 18.246) or academic inquiry (Plato, *Apology* 22e, *Theaetetus* 210c, *Philebus* 55c) rather than educational outcomes; and in Aristotle, an examiner, by this word, is an auditor of public accounts (*Politics* 1322b11). One

might describe an investigation through the term, but this was the scholarly activity itself, not the evaluation of someone's performance at a learning activity.

We have an adorable portrait of educational practice in ancient Greece in the dialogues of Plato. They represent Socrates talking to informal students who are keen to learn. It is never a process that ends with a discrete phase of assessment, though it could be held that it inherently assesses the higher virtues of discourse, as Socrates quizzes the fledgling philosophers to establish their analytical thinking. In their purpose, however, the point of the dialogues has nothing to do with assessing the neophytes. Throughout the dialogues, the interchange between the participants and master is to advance thinking rather than necessarily to establish the expertise of the student.

Learning in the ancient world, and in contemporary Indigenous cultures *par excellence,* was extensively informal. A potter apprenticed to a master was quite likely destined to become a master, but not on the basis of an exam. In Japanese ceramic tradition, students who have the privilege of studying with a great master will measure their success not by how quickly they can get through the prescribed activities but the reverse: how long they are accepted in the master's studio.

We do not know if there was assessment in ancient education; because philology does not tell us that something never happened just because it was not spoken about. Certainly, a boy had to qualify in some way to become an ephebe, for which the verb was to assay or test (δοκιμαζω, Aristophanes, *Wasps* 578; *cf.* Demosthenes 27.5); and a papyrus in London from the late Roman period does refer to the testing of recruits using this word (*Greek papyri in the BM* 3.982.6; *cf.* 2.338.24 from the second century using εξέτασις). But none of this conjures up an image of exam papers; it is more like the sporting trials (*athletae probatio*) in ancient Rome (Cicero, *De officiis* 1, 40, 144). When education features in the iconography of Greek vase painting, it looks like learning and seems too conversational for a process resembling assessment today. In language, the concept of examination was not only far from the delights of learning but is associated with stressful situations. The very word (εξέτασις) probably derives from a stretching (τεινω, hence τασις or swelling under force, as with the classical columns which show bulge, εντασις, perhaps related to the verb for stressing τανυω), which describes the pressure that you apply to something if you want to know that it is genuine and authentic.

Typically, this pressure might involve torture; for thus St Paul is to 'be examined by scourging (μαστιξιν ανεταζεσθαι)', *Acts* 22.24. It is hardly an educational situation; but it highlights one structural truth about assessment, namely that assessment is mainly for the benefit of the assessor, not the assessed. The same verbal root is used with Herod examining (επιζητησας) the guards, whom he subsequently puts to death (*Acts* 12.19). Examination in this context just means being tried or judged; indeed, in other places, the very torment that Paul is threatened with is about 'judging me (ανακριναντες με)', which occurred without Paul needing to be flogged (*Acts* 28.18).

Assessment as trial and trying

These stressful examinations are commonplace throughout western literature, where being examined may involve pain—on the assumption that greater truth can be extracted in a circumstance of exaggerated physical stress—and may also have fatal consequences. Examination is close to *trial*, which is a word by which we describe rigorous method in clinical and social science as well as law. When a person is tried in court, anxieties rise, because the

outcome may be grave: the person tries to defend himself or herself but is tried by another person or several on behalf of a state institution. The aim of a trial is to sift the truth from false impressions; and indeed the origins of the word trial are technical, relating to sifting, perhaps grinding and threshing (Oxford English Dictionary).

In common language, the idea of trying (try, try, try again) is a great mantra of encouragement. To get students to try, we build up their self-esteem and confidence and see the enterprise as affirming and reinforcing. But what is the next step that we ask them to do with their trying? It is to be tried, and all too often with catastrophic educational results (Boud, 1995, p. 43). The origins of this word are also on the gruelling side, as in Dryden's translation of Virgil's *Georgics*: 'To repair his strength he tries: / hardning his limbs with painful exercise' (3.107). To try, as in try on a shoe, is to see if something fits. It presupposes a template or a standard against which another item or person is measured. When I try something, either I or the something is, in a sense, on trial.

It is also no accident that to try in French (*essayer*) is also a genre of assessment as well as a literary or philosophical genre (*essai*). In writing a shorter disquisition or essay, one attempts or tries to broach a theme; but as this essay is itself measured by somebody's prior knowledge of the theme, the try is submitted to judgement. So again, the uncanny reflexiveness of assessment is installed in language: while you try, you are yourself being tried. Already the Greek word for trying (πειραζω) is both about trying and being tested, and from the earliest period of the epic (*Odyssey* 16.319, 23.114). And as if fulfilling this logical reflexiveness, St Paul exhorts: 'Examine yourselves (εαυτους πειραζετε), whether ye be in the faith; prove your own selves (εαυτους δοκιμαζετε)', which is effectively self-assessment (*2 Corinthians* 13.5).

Origins of examination

The Roman view of examination (*examen,* later the more abstract *examinatio*) is technical and weak, with very limited currency in Latin, where the idea did not transfer to education. The origins of the word are technical, relating to the beam of a balance or scale and inherently embody the concept of a standard and a metric. In English, the word was used in an educational context, meaning something like an exam, only since the seventeenth century (Oxford English Dictionary). Instances of the Latin root in the Romance derivatives during the renaissance and baroque are rare and also—at least in our reading so far—*not* educational. For example, Littré finds the first use of *examen* as an educational exam only in the twentieth century.

Meanwhile, several words were used in Latin to describe the sometimes painful interrogation of suspects, especially inquisition (*quaestio*) which is paradoxically the origin of our innocent word question but which is exceedingly unpleasant in the history of the Mediterranean, culminating in the persecution of heretics from the middle ages to the seventeenth century. As in Greek, the more innocent words around the concept of examination (like *scrutinium*) apply to what we would call research rather than education.

In a contemporary educational context, the word examination evokes ranks of school children or undergraduate students completing a paper in strictly invigilated circumstances, with a dire hush and disciplinary action taken for breaches of isolation from assistance in the form of contraband notes. The atmosphere is aligned with the etymology; the collective linguistic history of the words in even as simple a phrase as "there will be an essay question in the exam" conjures up torture, trial and inquisition. Typically, the examination terminates a period of study and is taken

as proof of a student's grasp of the syllabus. Though the study leading up to the examination may or may not yield long-term learning, the examination itself is not a learning experience but is solely conceived as proof of learning having taken place.

Origins of the test and testing

Insofar as exams are somewhat traumatic for the student, they seemed, at various times, to suit the purpose. In the European psyche, the way to credentialize anything is to subject it to stress. A metal, for example, might be proven to be whole or elementally irreducible by being tested under flame; and the vessel for this proof—an earthen pot or cupel (*testum*)—is the late Latin origin for our word 'test'. One contemporary argument for examinations is that they provide an environment where plagiarism is less possible (Carless, 2009); or to use ancient metaphor, the work is elementally irreducibly the student's own.

A test is close in meaning and sound to testimony and the academic testamur (we attest); but the semantic and phonological similarities are a coincidence. The words do not have a common origin. Testimony derives from the independent concept of a witness (*testis*). It would be attractive and convenient to Joughin's (2009) definition to believe that the educational testing process derived from bearing witness to an achievement; but in fact the origin is more technical and telling: the proof in question is fierce, gruelling and somehow metallurgical, which our Greek word already indicated (Xenophon, *Oeconomicus* 20.14).

Before contemplating the recent history of assessment, which is so much about proving things, we might examine one final way that language reveals the fateful alignment of experience, testing and proof. The three concepts are effectively bundled up in one Latin verb to test (*probare*) which is echoed in derivatives in Romance languages. To try, to taste, and even to experience: these words are united with the concept to prove. Proof has its origins in *probare*; and the Latin word did not lose this connotation in modern languages, as in the stressful terms of judging a murder (Boccaccio, *provato alla corte*, *Decameron* 3.7) in the fourteenth century. From the beginning of the renaissance, it can be used poetically as evidence, as with 'a more certain proof (*piú certa prova*, Petrarch, *Canzoniere* 119.16–19)' or ultimate proof (136.8); but it can also be used to express the idea of a challenge, as when the poet says to the personification of Love that he has to beat a new and marvellous *prova* (270.1–4). In the same century, the curious nun in Boccaccio who has an inkling that copulation must be the most amazing experience in the world formulates a plan to fornicate with the gardener Masetto in order to test if it really is so good (*provare se così è*, *Decameron* 3.1). With a liberal spirit shared by the author, the other nuns have equal interest to check out what kind of an animal a man may be (*provare che bestia fosse l'uomo*); each wanted to try (*provar volle*) in turn how Masetto knew how to mount, trying and trying again that sweetness (*provando e riprovando quella dolcezza*), a phrase that comes up in the next story as well.

The meaning of the same word shades off into experience in general, as with Guicciardini's awareness of honour which, he says 'I have experienced in myself (*provato in me medesimo*, Francesco Guicciardini, *Ricordi* 118). Only with difficulty could the phrase be translated as 'proving it in myself'. In such cases, which are also evident in French (La Bruyère, *Charactères* 2.1, 2.54, 6.9, 10.51), the conception is more related to experience. It makes us wonder about the close link between experience and proof, given that we generally believe something to be true and proven if we experience it. On the one hand, it is as if they are the same thing, a great empirical gridlock—remembering that the word empirical also has the Greek word 'to try'

locked up inside it—; but on the other hand, the closeness of the conceptions is much to be emulated in education, if the proof can be close to an authentic experience, as opposed to a rule or an abstract proposition learned by rote. So although we can interpret these ancient linguistic obscurities as harbingers of educational obscurities, we can also perceive a necessary alignment with how we see education and assessment at their best, according to constructivist methodology where experience, or the first-handedness of knowledge, is privileged over instructional passivity, subsequently assessed with an authoritarian monitored process.

Nevertheless, experience and proof are not identical; and the seventeenth-century author La Rochefoucault cautions about the fondness with which we accept verities when they belong to our own conspectus: 'what often prevents us from judging analyses that prove the falsity of virtues is that we believe too easily that they are genuine in ourselves' (*Maximes* 7); and 'nothing proves better that death is redoubtable than the pains that philosophers go to in order to persuade us that we ought to despise it' (12). But for much French literature in the baroque, the words (*prouver*, *éprouver*, *preuver*, differences which La Bruyère puts down to merely orthographic convention, *Charactères* 15.73) are used for the idea of moral test, as in Racine: 'I will make a final proof of my courage' (*Bérénice* 5.6, cf. *Mithridate* 4.4, *Alexandre* 1.2, 5.2) which must be uttered with grim relish.

The structural role of anxiety

English separates these ideas with distinct words, though there is still overlap and some logical continuity, given that proof is often experienced as an ordeal, a test of a scary kind as well as a kind of assessment in its own right. The linguistic history reveals a degree of confusion that inheres in the concept of assessment, which is both objective and anxious. One might add that in any case, assessment is somewhat equivocal in that one examines a person on the basis of what he or she says or does; one examines someone on something (τινα περι τινος) as Plato says (*Phaedrus* 258d) rather than directly examining the something, which would make it research. Assessment is not research but a ritual, which can already be established by the sociology of examinations.

Perhaps in reaction to the regimented disciplinarian air of examinations, with their stomach-withering dread of a mental-blank, educational culture has developed the nicer term of *assessment*. It is not exactly a euphemism but is gratefully image-free, somewhat clinical and abstract, without connotations of students marching in file to sit at small desks for three sweaty hours, with mean guards pacing up and down the aisles. But nor is it too soft either. Indeed the history of the word also has a certain judgmental gravity about it. Before it first came to be used in an educational context in the 1950s, assessment was long used in the evaluation of tax, a usage familiar to us still today with its link to penalties and temptations to cheat. The etymology is the verb to sit (*sedere*). In order to establish if goods, land or services should be subject to tax, a notary or minister must sit in judgement. Acting out this archaic motif, when we assess students, we ultimately bring the marks to a Board of Examiners or similar, where they are received with certain deliberations by people who *sit* on the board.

The 1989 edition of the *Oxford English Dictionary*, of which the online version dates from 2010, still gives little recognition to the educational uses of the noun assessment. It features in one of five categories, as a subset of 'estimation, evaluation'; but as for the verb 'to assess', not even one educational usage is cited. In English before the second world war, assessment is not a very common word. We do not see any form of the word in Shakespeare, for example, perhaps

because the bard does not use the vocabulary of taxation. But if we check out the post-war period and we look to the great dictionary of the internet, the word assess achieves a Google search of 157 million pages; with many focused on educational uses of the term (July 2012). The noun assessment gets 83 million, and Google's suggestion of 'assessment for learning' yields almost 19 million hits.

Modern educational uses of the term assessment in the scholarly literature are almost entirely post-war; articles in ERIC do not use the term in titles before 1959, and much of the early literature is dominated by the assessment of teaching; this may be the source of Northern American meanings of assessment as a synonym for the evaluation of educational methods or systems, rather than assessment of and for the individual. The term is used around 20,000 times in Google Scholar before 1945, however the uses most compatible with Joughin's definition relate to psychometric and physical testing.

Coming from an asymptote approaching zero, the growth in assessment is exponential and we now appear to be in a global epidemic of assessment language.

Etymology guides purpose?

Assessment in the contemporary academy performs a variety of purposes that extend it beyond the scope of the term 'evaluation'. Assessment is at once a summative credentializing, as well as a formative guide towards learning; assessment is focused on an immediate task, as well as inculcating the values of a field of study. Boud (2000) describes this simultaneity as 'double duty'. To view the terms 'assessment' and 'evaluation' as synonymous is to lose assessment's other purposes. Some purposes, roles or goals of assessment are stated explicitly, others assumed or unstated. Of the myriad purposes that assessment holds, which are encapsulated in the history of the language of assessment?

We have historicized the language of assessment, and its history may be connected to unenlightened default assumptions about the purposes of assessment. Notions of a rigorous stressful challenge with a goal of attainment have philological foundation. It would appear that modern notions of assessment privileging learning have less of a foundation in history; interestingly when university policywriters attempt to define assessment, they are more similar to the ancient than the modern (Boud, 2007). Perhaps the linguistic baggage of terms like assessment, essay, test, try, exam and prove might be what is clouding our thinking about what assessment could be, and keeping policywriters attached to what it was.

Reclaiming the language: Assessment as learning

Is it doing us any good? As noted, all processes of testing and examination are about proof and, in being put to the point, one could be proved awfully lacking. Assessment and anxiety go together. We remember the palpitations of being examined when we were students and academics remain scarred by the experience, as Boud notes: 'even successful, able and committed students—those who become university teachers—have been hurt by their experiences of assessment, time and time again, through school and through higher education' (Boud, 1995, p. 35).

In recent times, academics have also expressed their anxieties about pressures on research performance forcing teachers to make odious efficiencies in their assessment. This fear is so strong that it has moved one university executive to voice scruples that the student experience is imperiled by the Muliple Choice Question (MCQ) method (Bailes, 2012). Indeed, to determine the method of assessment on the basis of expedience offends all educational principle; but there are grounds for heeding the pessimism. The advent of the Learning Management System (LMS) has made MCQs efficient and readily available to an unprecedented degree; and the ubiquity of LMSs presages a boom in automated assessment, which seamlessly enters marks in a gradebook and saves hours of work for academics. There is no need to read anything, nor even to tally the marks, because the computer calculates them infallibly.

Concerns have been expressed for the quality of the educational process as well as student experience. If assessment defines the actual curriculum (Ramsden, 1992), then curricula built around MCQ testing privilege short-term factual recall of truths and identification of falsehoods. The MCQ approach can even lead to learning of 'false facts' present in distractor options when feedback is not immediate (Butler & Roediger, 2008; Marsh, Roediger, Bjork, & Bjork, 2007). It would appear from these observations that a multiple-choice questionnaire in an LMS is an unlikely place in which to solve the increasing difficulties of mechanical assessment for the sake of the assessor rather than the student's learning.

This is the context in which we propose a further development in the history of assessment. Along with Boud (2000), we believe that assessment is inferior if it is not itself an educational experience. Of course it is easy enough to establish this preference in principle; but in practice, the need to invent assessments that are simultaneously a learning experience adds to the lecturer's stress of creating effective, accurate, undiscriminating and comprehensive assessment.

Working toward assessment-as-learning, our innovation is to use the multiple-choice structure in any LMS to perform the opposite of a multiple-choice test. We have reconfigured the MCQ apparatus so that it becomes the host to a Socratic dialogue of a speculative open-ended conversational character rather than a quiz of a finite and absolute nature. This system, which we call a conversation simulator, still delivers assessment but in a way that is integral with learning.

In a conversation simulation, the computer presents the student with a situation, perhaps a problem which seems complicated and difficult to answer. A solution is then proposed, to which the student is asked either to agree, to disagree or neither to agree nor disagree: yes, no or maybe. Based on the choice that the student makes of these options, the computer then provides feedback on the wisdom of the choice and thus entertains the student in further reflection. These interactive encounters with the conversational text are (a) a learning experience with the mixture of narrative and argument that is proper to a tutorial and (b) assessable according to exact and consistent criteria with a numerical expression. And because the conversation simulator is built into the LMS, it scales up effortlessly and handles any number of students.

A conversation simulation functions according to the following sequence

    problem → proposal → decision → reflexion

the theory of which is explained in a comprehensive open-source manual by the authors (Nelson & Dawson, 2013). Consider the following excerpt from a postgraduate unit for doctoral supervisors:

A problem

Two researchers in social medicine have devised a plan to investigate the hidden milieu of online anorexic communities. They are extremely secretive and members on pro-anorexic sites are suspicious and exclude all forms of research. One of the investigators adopts a pseudonym, uses the language of youth and projects all the neuroses to gain acceptance. How ethical is this methodology?

A response

It sounds ugly but we have to remember that anorexia is a serious condition, akin to suicide, and unless we understand how it is handled, we cannot advance medical science.

Do you agree with this response?

No, Maybe, or Yes

Feedback for reflexion

The preferred answer is Maybe. The response doesn't answer the ethical question. It's true that we want to understand anorexia; but does that mean that we have to resort to deception? The investigators are conducting themselves in a somewhat fraudulent spirit.

(Problem posed in HED5070: Postgraduate research supervision, Graduate Certificate of Academic Practice, where the same problem is presented four times, on each occasion with a different response to search out the ethical rationale of various research methodologies.)

Conversation simulations, as well as functioning as assessment, are an educational experience. As with Socratic dialogue, the conversations perform the teaching; if MCQs "modify the knowledge they are designed to assess" (Marsh et al., 2007, p. 197) then rather than attempt to limit this effect we embrace it. Because teaching and assessment are indivisibly folded into one another, there is no grey syllabus, no teaching of A, B, C and D but only assessing B. In that respect, assessment is anxiety-free, because the teacher has no fear of redundant syllabus. On the student's side, too, there is little anxiety, because the learning activity automatically clocks up credit, and this process is experienced as intellectually stimulating.

Is it possible that a kind of *deus ex machina* can deliver assessment from its escalating anxious patterns, which rehearse too many of the bad memories of history? We will not be in a position to know the true efficacy of the conversational simulator until it has been more broadly adopted (and tried, examined and tested by flame) by other educators; however, our own experience with this innovation leads us to cautious optimism. The computer has facilitated a stress-free neo-Socratic encounter with ideas. In our emergent mass and massive higher education future, anxiety around education matters even more (Nelson & Dawson, 2012), and perhaps this or other methods of Socratic dialogue might help ease the anxiety. This technical discovery encourages us in thinking that the historically rising anxiety around assessment can be broken, by leaving its linguistic baggage behind.

Notes on contributors

Robert Nelson is Associate Professor and Associate Director (Student Experience) at Monash University. The key theme of Robert's research is how the aesthetic interacts with the moral and the educational. Phillip Dawson is a Lecturer in the Office of the Pro Vice-Chancellor (Learning

and Teaching) at Monash University. Phillip's research explores how academics make decisions. Email: robert.nelson@monash.edu

References

Numbers to classical and biblical texts follow the standard reference system derived from the canonical editions used in lexicography. Different editions and translations have disparate pagination, whereas the canonical numbering is consistent. Similarly, the numbering used in renaissance and baroque texts follows the order of canto/stanza for epic poems, act/scene for plays or book/story for *novelle* or essays.

Bailes, M. (2012, June 13). Strength in numbers: do ERA rankings add up for universities?, *The Conversation*. Retrieved from http://theconversation.com/strength-in-numbers-do-era-rankings-add-up-for-universities-7493

Boud, D. (1995). Assessment and learning: contradictory or complementary. In P. Knight (Ed.), *Assessment for Learning in Higher Education* (pp. 35-48). London: Kogan Page.

Boud, D. (2000). Sustainable Assessment: Rethinking assessment for the learning society. *Studies in Continuing Education, 22*(2), 151-167. doi: 10.1080/713695728

Boud, D. (2007). Reframing assessment as if learning were important. In D. Boud & N. Falchikov (Eds.), *Rethinking Assessment in Higher Education: Learning for the longer term* (pp. 14-28). London: Routledge.

Butler, A., & Roediger, H. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*(3), 604-616. doi: 10.3758/mc.36.3.604

Carless, D. (2009). Trust, distrust and their impact on assessment reform. *Assessment & Evaluation in Higher Education, 34*(1), 79-89. doi: 10.1080/02602930801895786

Dawson, P., Bearman, M., Boud, D. J., Hall, M., Molloy, E. K., Bennett, S., & Gordon, J. (2013). Assessment Might Dictate the Curriculum, But What Dictates Assessment? *Teaching & Learning Inquiry: The ISSOTL Journal, 1*(1), 107-111. doi: 10.2979/teachlearninqu.1.1.107

Joughin, G. (2009). Assessment, Learning and Judgement in Higher Education: A Critical Review. In G. Joughin (Ed.), *Assessment, Learning and Judgement in Higher Education* (pp. 13-27). Dordrecht: Springer Netherlands.

Marsh, E., Roediger, H., Bjork, R., & Bjork, E. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review, 14*(2), 194-199. doi: 10.3758/bf03194051

Nelson, R., & Dawson, P. (2012, October 9). MOOCs and exercise bikes – more in common than you'd think, *The Conversation*. Retrieved from https://theconversation.edu.au/moocs-and-exercise-bikes-more-in-common-than-youd-think-9726

Nelson, R., & Dawson, P. (2013). Conversation Sim Manual Retrieved 22 March, 2013, from http://www.conversationsim.org

Oxford English Dictionary. *"examination, n."*: Oxford University Press.

Oxford English Dictionary. *"try, v."*: Oxford University Press.

Ramsden, P. (1992). *Learning to teach in higher education*. London: Routledge.