Article

e-learning and
digital media

E-Learning and Digital Media
2017, Vol. 14(3) 123–137

DOI: 10.1177/2042753017732503
journals.sagepub.com/home/ldm
SAGE

Economizing education: Assessment algorithms and calculative agencies

Cormac O'Keeffe
Lancaster University, UK

Abstract

International Large Scale Assessments have been producing data about educational attainment for over 60 years. More recently however, these assessments as tests have become digitally and computationally complex and increasingly rely on the calculative work performed by algorithms. In this article I first consider the coordination of relations between the human and non-human agents that perform the day-to-day tasks of data production used in economic and educational policymaking and practice. I examine the calculative agencies of an assemblage of algorithms encoded in the testing software for the Programme for the International Assessment of Adult Competencies. These algorithms perform the sampling, sorting, scoring, and result prediction of test takers and items during digital assessment events. Second, I examine the role of psychometric practices and educational testing theories, and in particular, Item Response Theory, in the work of sorting and detaching situated practices into equivalence spaces that they can be manipulated and transformed by into calculable entities. Combined with digital assessment technologies, the probabilistic statistical techniques used by Item Response Theory are able to produce digital data such as test scores capable of transforming situated literacy practices into psychological constructs that can then be classified and rendered calculable. This reinforces the calculative agency of tests as well as a consensus about the legitimacy and necessity of the testing technologies as the dominant way to produce educational data.

Keywords

Testing, algorithms, data, economics, International Large Scale Assessment, Item Response Theory, STS

Introduction: Big tests, big business, and governance

In August 2016, representatives from French corporations, the European Commission, and the French Education Ministry met to discuss how to strengthen ties between education and business. Jean-Marc Huart, the French government representative, declared that

Corresponding author:
Cormac O'Keeffe, YES, 40ter avenue de suffren, Paris 75015, France.
Email: okeeffe.cormac@gmail.com

'without businesses, education cannot do anything' (Lambert and Leder, 2016: 3). This rapprochement between business and what have traditionally been state-run education and training sectors is not a recent phenomenon nor by any means recent or limited to France. In 2004, while serving as Chief Adviser to Education for Tony Blair before later becoming Chief Education Advisor to Pearson, Michael Barber, enthusiastically predicted this transformation stating that 'just as financial services globalized in the 1980s, and media and communications in the 1990s, so in the, 2000 [sic] we will see education reform globalising' (Barber, 2004: 40). The accuracy of this prediction has been evidenced by the process of globalization where knowledge 'is treated as a fictitious commodity' (Ozga et al., 2006: 8). This has been accompanied by narratives that challenge the concept of education as a right (Luke, 1997) and reframe and redefine education 'as a sector for investment and profit-making' (Verger et al., 2016: 3). This trend towards commodification is reinforced by the human capital hypothesis where 'the progress of technological rationality is supposed to lead automatically to the triumph of human capital' (Piketty, 2014: 27) or skills development in 'knowledge-based economies' (OECD, 2013a: 3), thus melding edu-business (Ball, 2007; Lingard and Sellar, 2013) with the public good. Although a correlation between an increase in basic skill development (such as literacy and numeracy) and rising wealth for all is far from having being proven, this 'literacy myth' (Graff, 2012) is a powerful one that shapes public imaginaries (Hamilton, 2012).

In order to make education and training amenable to policymaking and business plans, the everyday practices of schools, universities, and training centres need to be translated into calculable data. These data are particularly amenable to business and governmental practices that value ranking and prediction. The intensely utilitarian character of these data production practices also ties into narratives that support a development agenda that advances a particular 'economistic view of educational aims linked to the requirements of a global knowledge economy and ideas about educational governance linked to new public management, which increasingly promote corporatized and privatised administration of education, outcome measures and knowledge as commodity' (Lingard and Grek, 2005: 8–9). The intensification of digital data production has meant that many domains of activity have increasingly come under scrutiny as marketable (Fourcade and Healy, 2016) and the growth of algorithms for 'harvesting personal data and making it economically productive' (Mager, 2016: 3) are becoming indissociable from modern business practices.

Tests, and in particular the increasingly common International Large Scale Assessments (ILSAs) produced by international organizations (IOs) such as the The Organisation for Economic Co-operation and Development (OECD), the International Association for the Evaluation of Educational Achievement (IEA), European Commission, and UNESCO along with private companies such as Educational Testing Services (ETS), and Pearson, are especially efficient types of 'knowledge production apparatus[es]' (Morgan, 2016: 10), for the production of data in the global marketplace of the knowledge economy. The business of educational data production (Hogan et al., 2015a) is tightly dependent on the statistical sciences, especially psychometrics, that produce data about educational attainment (Williamson, 2016). Tests are key technologies since they allow proponents of human capital theory (Hanushek et al., 2013; OECD, 2013a) to argue that 'educational performances and outcomes can be scientifically quantified, standardised and normed, and thereby educational processes and practices can be improved and made more efficient in the production of laboring power' (Luke, 1997: 3).

Although there is debate as to whether or not knowledge and ideas can be properly considered as a type of capital at all (Dean and Kretschmer, 2007; Luke, 1997; Piketty, 2014), I suggest that that knowledge is being framed as a commodity and the practices that produce it as a market. It is outside the scope of this paper to argue one way or another in favour or against the presence of private or public interests in education or testing. Rather, it is to highlight the extent to which educational testing does more than give data to the market, but how it is algorithmically configured to marketize data production. Testing events transform literacy and numeracy practices into test data on skills that can be framed as commodity or good as, 'its properties represent value for the buyer' (Callon and Muniesa, 2005: 1233). The attribution of exchange value to the specific forms of knowledge produced by tests moves test data into a 'commodity state' (Appadurai, 1994: 83) so that the practices of education can be exchanged between educational, business, policy, and testing cultures that each have their 'bounded and localized system of meanings' (p. 84).

In the paper that follows, I use the theoretical framework of algorithmic configurations proposed by Callon and Muniesa (2003). Originally used to analyse economic transactions and markets, I use it to explore how digital assessment technologies, with a particular focus on algorithms, organize and frame the socio-material calculative agencies that perform the work of computer-adaptive testing (CAT). I describe how statistical technologies are deployed during digital testing events and analyse how the calculative agencies of an assemblage of actors that are involved in the work of sorting and sampling, classifying, ranking, and predicting during digital assessment events play an important role in the production of educational data.

In particular, I focus on the Programme for the International Assessment of Adult Competencies (PIAAC). Run by the OECD since 2008, this digital assessment is a psychometric instrument with the purpose of producing data on the 'skills' or 'competencies' present in their adult populations (16–60) much in the same way that its most famous product, the Programme for International Student Assessment claims to do for school-age children (OECD, 2016). These data are paid for by national governments. Its for-profit extension, Education & Skills Online, targets, among others, companies 'that want to use the results to help them identify the training needs related to literacy and numeracy for their workforce' (OECD, 2015: 1). These are examples of tests that exemplify many of the characteristics of the data production technologies that are able to singularize the concept of human ability. That is, they are able to detach situated practices from their context and make them into calculable goods for use in policy and economic decision-making. Tests are particularly apt at this kind of framing as they are able to draw a 'boundary between goods included in the space of market calculation and those that were excluded' (Callon and Muniesa, 2003: 1235) through the enactment of circumscribed assessment events.

Finally, I examine how the complex relations between the human and non-human calculative agencies that perform the work of CAT are characterized by chains of action that are tightly scripted yet as situated and contingent as the literacy practices they assess. The ensemble of data production practices that constitute the testing industry generate 'authoritative knowledge' (Brun-Cottan et al., 1991) that is used by policymakers, IOs, states (Grek, 2014; Williamson, 2015), economists (O'Keeffe, 2016), and private and quasi-private business networks such as ETS and Pearson (Ball and Junemann, 2012; Hogan, 2015b).

PIAAC and its calculative agencies

PIAAC is at once an old and new test. Old, in that it inherits many of the items, workflows, algorithms, psychological, statistical technologies, and political assumptions that come from older tests such as the Adult Literacy and Lifeskills Survey (ALL) and the International Adult Literacy Survey (IALS). New, since PIAAC marks a genuine change in that it was the first time that an International Large-Scale Assessment (ILSA) was almost entirely digital (Thorn, 2014). PIAAC is a pertinent object of focus for exploration into algorithmic culture and education since 'many features found in the current PIAAC foreshadow the future of ILSA' (Yamamoto, 2011: 10) and the production of digital data on educational attainment.

PIAAC, like most ILSAs, uses a psychometric technique known as Item Response Theory (IRT). For any educational test that uses IRT, its primary purpose is to determine the value of theta (θ) or a test taker's ability. Ability is understood as a characteristic that is not directly observable, otherwise known as a latent trait. IRT has a number of core assumptions one of which is the notion that practices such as literacy can be defined as a unidimensional construct (ability). As such, for test-makers, it can be 'assumed that, whatever the ability, it can be measured on a scale having a midpoint of zero, a unit of measurement of one, and a range from negative infinity to positive infinity' (Baker, 2001: 5). Furthermore, people and items can be measured and placed on the same scale. IRT-based tests use these assumptions to create what are known as item characteristic curves. With these curves, psychometricians can create models that are able to plot the probability of a correct response to an item on the ability scale and predict a test taker's ability to respond correctly to an item of known difficulty even when no data other than imputed values exist.

IRT is at the heart of, and inseparable from the algorithmic design of PIAAC's testing workflow. It was once one of a number of competing approaches employed in machine learning and digital assessment that is used to make claims about abilities or skills. However, over the last 60 years, IRT has come to assume a dominant position in the area of educational testing theories and practice having vanquished many alternatives such as Classical Test Theory (CTT) (Goldstein and Wood, 1999; McNamara and Knoch, 2012). Despite being more mathematically complex and requiring larger sample sizes than CTT, IRT has allowed ILSAs to create ability models that are independent of both specific populations or tests and thus much more amenable to the production of large-scale, international comparative statistics. The psychometric practices behind the creation and delivery of test items are so tightly woven into the mathematics, history, and politics of large-scale literacy and numeracy assessment (Carlson and von Davier, 2013; Goldstein and Wood, 1999) that much of the legitimacy of the test and the organizations that produce it hinge on the performance of IRT and its many calculative agencies.

As IRT is primarily a probabilistic approach to assessment, it requires a great deal of computational effort to work effectively, particularly when determining item parameters (such as difficulty). Since much of this calculative work is given over to software and computer-based actors, IRT increasingly relies on a number of algorithms to be performed (Wang et al., 2010: 21).

This study will look in detail at a computer-based adaptive assessment algorithm developed for PIAAC that was of many in a long chain of inscriptions, frequently algorithmic in both design and execution that produced data on adult skills.

Methodology: Human and non-human calculative agencies

This paper adopts a material semiotic approach to investigating algorithms that draws on the work done in the anthropology of science and techniques. This approach problematizes a number of framings of algorithms as neutral tools, natural occurrences, anthropomorphic entities, or even quasi-magical entities. A common depiction of algorithms is that they are nothing more 'a sequence of instructions telling a computer what to do' (Domingos, 2015: 4). Such a definition, although succinct, makes a number of assumptions, the most of important of which, is that an algorithm is little more than delegation of human agency to code. Algorithms are not passive or neutral objects that spring into action once activated. Although for the most part, the agency of algorithm are tightly scripted and predictable, their actions can also have unintended consequences. For instance, 'due to an erroneous selection algorithm at the second stage of sample selection in PIAAC Germany, person probabilities of selection were unequal, so that some persons had a greater chance of being included in the sample than others' (Zabal et al., 2014: 82). While this is far removed from the 'overly dramatized' and apocalyptic vision of High Frequency Trading algorithms capable of making markets crash and corporations crumble within milliseconds (Bondo Hansen, 2015), it takes us away from the somewhat naive idea that the precision with which an algorithm is written remains the same once it is out in the wild and interacting with other agents.

At the opposite extreme of this portrayal of algorithms as neutral tools, is depicting them as anthropomorphic or natural entities with their own innate qualities and characteristics. The 'danger to be guarded against here is taking an essentializing view of algorithms' (Dourish, 2016: 2) that masks their emergent, and often messy characteristics as parts of wider sense-making assemblages (Neyland, 2015). One of the difficulties with algorithms is that their often inscrutable behaviour to non-specialists means that they remain black-boxed (Pasquale, 2015) – unintentionally or otherwise. The mysterious algorithmic character of their operations can make them appear quasi-magical and thus inscribes them in a long tradition of 'mathematics and number magic' (Goody, 1975: 17) where the algorithmically literate are like priests with grimoires in preliterate societies. Coders and mathematicians, who, with the manipulation of letters and numbers, are able to be attributed with a kind of modern-day 'sourcery' (Chun cited in Ziewitz (2016: 5)).

Instead, it is more helpful to focus on not so much what an algorithm is, but on what it does. There is a great deal to be learnt by investigating how algorithms, as socio-material entities actively participate in creating the chains of negotiations, mistakes, discussions, and happenstance that characterize any endeavour. Algorithms, like the computers that house them, are 'both material, in the sense that they both are things that "hold together" and that can be appropriated because they have objectified properties' (Callon and Muniesa, 2005: 1233). That is, algorithms, while lacking the physicality of a computer or server farm, go through a processes of materialization in that they are 'given stable forms as objects and as categories and distributed between processes, actors and interests' (Slater, 2002: 109).

Algorithms, like any other entity, 'rely on materially existing infrastructures, artefacts, people, and practices – often operating behind the scenes – that often fundamentally shape and structure how those seemingly immaterial abstractions operate' (Geiger, 2014: 347). The cognitive effort, and in the case of ILSAs, the extremely complicated, expensive and time-consuming effort of calculating ability, must be distributed among an array of material entities. This material semiotic approach to studying entities and the relations between them develops the notion that that communication is not understood as exclusively between human beings but between many actors (or actants) distributed along a continuum of materiality since, 'knowledge and action are never individual; they mobilize entities, humans and non-humans, who participate in the enterprises of knowledge or in action. This participation is active and can only exceptionally be reduced to a purely instrumental dimension' (Callon, 2005: 1237).

Algorithms are understood as relational entities and the calculative agencies of algorithms in PIAAC are distributed and involve both humans and non-humans (Hutchins, 1995) involved in the assessment process. Seen together, the assemblage of actions that make assessment practices form what can be termed algorithmic configurations.

Originally developed for studies of markets, Callon and Muniesa's (2003) model of algorithmic configurations is particularly amenable to the business of educational testing. My argument is that the operationalization of knowledge and skills as observable and testable entities such literacy and numeracy allows them to be singularized as commodities with exchange value in the data marketplace of edu-business, IOs, and governments. One of the roles of ILSAs in a knowledge society is to transform the outputs of education, notably literacy or numeracy and by 'imagining it as a crucial commodity within the knowledge society, a high-value positional good that can be cultivated and exchanged to the mutual benefit of all' (Hamilton, 2012: 27). Algorithmic configurations are calculative devices that,

> a) circumscribe the group of calculative agencies that are to be met, by making them identifiable and enumerable; b) organize their encounter, that is, their connection; and c) establish the rules or conventions that set the order in which these connections must be treated and taken into account. (Callon and Muniesa, 2003: 27)

In the section below, I examine the algorithmic configurations that calculate ability during the assessment events of PIAAC.

What do the algorithms in PIAAC do?

PIAAC is much like any other complex assemblage of testing software in that it relies on a range of processes and algorithms to perform the production of educational data. Each algorithm contributes to the overall success of the assessment event. In the section below, I will focus in detail on the work done by certain non-human actors. That done by the human actors, and how the agencies become entangled is described in greater detail elsewhere (O'Keeffe, 2015, 2016).

The calibration, delivery, and analysis of test items and results are complicated and expensive processes and PIAAC's algorithms automate a range of mathematical and statistical procedures (see Table 1). Some, such as the implementation of the Longest Common Subsequence algorithm are designed to mimic the 'leniency' of human test scorers (OECD, 2013b). Others are designed to control and monitor human error that occurs during the background questionnaire phase (OECD, 2011) while others are designed to replace human behaviour and optimize post-test analysis by speeding up processes that would normally take much longer to accomplish (OECD, 2013b). Different algorithms, such as the screener algorithm that ensured that sampling distributions are correctly implemented, allow human agency to be exerted at a distance. PIAAC's algorithms are also networked and their calculations are reported back to testing centres that both coordinate their work and that of the human interviewers (Hogan et al., 2016).

Table 1. A selection of algorithms and their actions within PIAAC.

Phase	Algorithm	What does it do?
Before testing	Cox's controlled rounding algorithm	This algorithm was used to ensure that the distribution of selected municipalities in the sample reflected the distribution in the whole population.
	Sampling algorithm	'A sampling algorithm for the screener to determine who, if anyone, in the household is eligible to participate in the study' (NCES, 2013: 6)
During testing	Adaptive algorithm	An algorithm to reduce response burden by calculating individual person characteristics, basic cognitive skills, and assigning them to one of 12 possible assessment paths.
	Levenshtein	This algorithm tracked all additions, suppressions, and modifications in the assessment process using a 'diffing' technique that compared different versions of reports.
	Longest common subsequence algorithm	This algorithm allowed the software to score items where the test taker had to highlight text. It was designed to mimic the leniency of a human scorer by tolerating incomplete but correct responses.
After testing	Expectation-maximization algorithm	A two-step algorithm commonly used in IRT that attempts to compensate for the insufficient number of test items needed to observe a latent trait (ability/theta θ)). The first step (expectation) estimates the likelihood (probability) of an examinee's response pattern to test items. That is, what one would expect the person to have responded. The second step (maximization) takes the unknown parameters estimated by maximum likelihood by treating the estimated sufficient statistics as observed.
	Plausible values algorithm	This algorithm uses the latent regression IRT model as an imputation model to generate different ability measures (in the case of PIAAC, 10) to compensate for the fact that the test taker did not see enough items to adequately assess the construct being measured.
	Newton–Raphson algorithm	Used to calculate the maximum likelihood of ability/theta (θ) for skill use at work (as part of the background questionnaire.
	Classification tree algorithm	A nonresponse bias analysis (NRBA) process where the algorithm selects subgroups within populations by estimating potential low or non-response rates.

IRT: Item Response Theory; PIAAC: Programme for the International Assessment of Adult Competencies.

The PIAAC assessment event's adaptive algorithm

For PIAAC, each test taker was given a laptop and that allowed them to interact with the TAO delivery system – the software that runs the test (Figure 1). PIAAC's for-profit version, Education and Skills Online is only slightly different in that it uses a web service through the test taker's own computer. Nested within the TAOs software was a probability-based multistage adaptive algorithm. As one of the core algorithms in PIAAC it also enacted IRT and allowed PIAAC to ensure that,

> task difficulty was adapted to a respondent's individual characteristics. Thus, the difficulty of the assessment items varied, depending on the information derived from the background questionnaire and the performance in previous parts of the computer branch. Adaptive testing enabled a deeper and more accurate assessment of respondents' ability level, while reducing respondents' burdens. (Zabal et al., 2014: 17)

Drawing upon a possible 166 items from an item bank, each test taker was moved down one of 12 different paths. The algorithm was able to measure a skill pattern by assessing ' multiple domains in an adaptive manner using background data, country-level prior information from previous assessments, and a stage 0 pretest of a limited number of core items in multiple domains' (Von Davier and Ying, 2016: 226). Expressed as an equation, the algorithm

$$B_{BQ} + \text{Core} = \arg\min\Big(E\Big(-\ln\Big(P_{\text{post}}\Big(\widehat{\alpha}_{\text{V}}^{(\text{BQ+Core})}\Big)\Big)\Big)\Big)$$

sorted learners ($_{\text{v}}$) into different initial categories based on data from the background questionnaire (B_{BQ}) and questions from a simple pretest (Core). Of the 455 possible questions in the background questionnaires, two were chosen as proxy cognitive indicators to categorize test takers into one of five possible categories along a continuum ranging from low educational attainment and no native language proficiency to high educational attainment and native speaker proficiency: 'not finished high school and nonnative speakers (low/no), not finished high school but native speakers (low/yes), finished high school but non-native speakers (mid/no), finished high school and native speakers (mid/yes), and have higher education (high)' (Chen et al., 2016: 399).

Here, the algorithm classified test takers and it is only once they have been sorted out that they can be taken into account 'associated with one another and subjected to the manipulations and transformations' (Callon and Muniesa, 2003: 1231) necessary for their embodied knowledge to be transformed into data for international comparisons of human capital or skills.

The work of item calibration was performed by another algorithm that worked in tandem with the adaptive algorithm and 'monitored DIF [Differential Item Functioning] measures and that automatically generated a suggested list of country specific item treatments' (OECD, 2013b, Chapter 17, p. 5) grouping and assigning items depending on the country of origin of the test taker. Once this had been done the test taker's ability was measured and plotted along a skill vector producing the test taker's latent cognitive profile ($\widehat{\alpha}_{\text{V}}$).[1] The skill vector was constantly measured and calculated. This was done by calculating Shannon entropy ($E(-\text{ln}$), that is the amount of information given by the test, to ensure the minimum number of items for each block (arg min) were correctly given. This was to ensure that the test taker would have best probability of receiving items at an appropriate level of difficulty and ensuring that this 'item exposure' rate was correct, that is giving the respondent just enough of the optimally calibrated items to measure their ability. These calculations were updated by how the test taker interacted with each item, that is the algorithm applied a posterior score distribution to the skill vector (P_{post}).

While all of this was happening, the adaptive testing algorithm was being given data produced by the LCS algorithm that controlled some instances of the automated marking (see Table 1). The Levenshtein algorithm constantly monitored the assessment process logging all 'additions, suppressions and modifications performed on an assessment process' before sending the data back to the relevant national database and allowing this process to be mapped.

The scoring procedure could only be enacted by the software and hardware present during the assessment event since no human could possibly have enacted this assemblage of algorithms with either the same speed or precision. In this respect, the speed of the adaptive testing algorithm embodies:

> forms of cognition, specific styles of perception that, on the one side, are non-anthropomorphic in the sense that they consist of procedures that can only be enacted by fast computing machinery, and, on the other side, are thoroughly entangled with operational arrangements. (Rieder, 2016: 30)

The adaptive testing algorithm developed for PIAAC offered the possibility of optimizing test delivery to the point that the significant investment of PIAAC digital and redistributing the work among encoded actors meant that PIAAC was made into a digital assessment as much for the algorithm was made for the computerized test of PIAAC. The calculative strength of the algorithm within this testing assemblage is significant since assessment events become dominated by relations of calculation (Callon and Muniesa, 2005).

Discussion

Admittedly, the details of these algorithms are complicated and the knowledge and experience required to adequately engage with them means that much of the process of testing 'itself is too technical for most to understand. Indeed, such calculations are not only inaccessible to non-experts in terms of comprehending them, but also in challenging them' (Gorur and Koyama, 2013: 636). In this brief description, my intention is not to debate the mathematical merits of the algorithms themselves. Nevertheless, it is important to analyse the methods used to produce educational data since 'educational assessment, perhaps more than any other aspect of education, has suffered the thraldom of "methodological empiricism" in which questions of technique have effectively predominated over the more fundamental issue of its effects' (Broadfoot, 2011: 118).

The adaptive algorithm played a role in circumscribing a range of judgements about what is worth being considered, or measured, during assessment events and shows how assumptions about human behaviour and life experience are folded into the enactment of the assessment event. In the example of the above algorithm, as schooling is taken as a proxy for cognitive ability, someone who never completed high school will never have the same possibility to answer difficult questions as someone with a university degree – the item response model that informs the algorithm will not allow this.

PIAAC is especially adept at this kind of boundary setting. A fundamental methodological principle in IRT is the notion of 'conditional independence (sometimes also called local item parameters)' that states that 'there is no dependence on any demographic characteristics of the examinees, or responses to any other items presented in a test, or the survey administration conditions' (OECD, 2013b, Chapter 17, p. 3). IRT thus attempts to extract itself from its context (while choosing which contextual elements to include in its calculations). Although tests are as situated and contingent as the practices they assess, this approach makes each instance of a test a 'mini-laboratory for the inscription of difference, enabling the realisation of almost any psychological scheme for differentiating individuals in a brief time, a manageable space, at the will of the expert' (Rose, 1996: 74). As such, tests lend themselves to the sorting and detaching of situated practices into equivalence spaces (Desrosiéres, 1993) where they can be manipulated and transformed by psychometric practices into models that lend themselves to being acted upon by calculative devices. The algorithm is adaptive but exclusively in terms of the item response model.

For this reason, an alternative name for IRT is Item Response Modelling. Goldstein and Wood (1999) argue that IRT does not attempt to explain

> why an individual should get an item right or wrong, or what conditions should be present for a particular outcome to happen, but about the supposed probabilistic nature of item response conditioned on something called ability. When person responses are modelled by a response function, the model *is* the theory. (p. 1)

PIAAC, like most tests, cannot find out what skills people have other than those that the test maker expected to find when designing the test. Before the first test taker interacts with the first item, the level of difficulty and the expected answers have already been decided upon long in advance and in some instances several years before the assessment event itself.[2] What the test and its algorithms can do, however, is to determine the quantities of people, at any given time, in any given place, align themselves with the item response model.

Although tests are good at producing educational data, one danger of tests is that they are very good at producing data about tests themselves. The algorithms used in the construction of PIAAC are there to measure ability and as such have a tendency to focus on 'estimating items and ability parameters and leave it at that, a tendency encouraged by computerised adaptive testing' (Goldstein and Wood, 1999: 163). In the words of an ILSA official:

> As a method to test an instrument it's as good as any other so why not? The big problem that we had with the OECD [excerpt removed to protect confidentiality] was the tool that they had, that they used to measure competencies was not up for debate and the tests were there for one purpose only: correct the coefficients of questions to estimate the scores for skills...So, the tool that they developed is fun, and it's what they wanted but it would never let us to challenge the tool itself and for us this was a major concern. (O'Keeffe, 2015)

Educational testing produces ever longer metrological chains of reference and the various actions (sampling, testing, analysing, reporting) all have the 'goal and the effect of *reducing the distance,* little by little, between the knowing subject and the known object' (Latour, 2013: 448). They thus create economic indicators that although they may be useful for calculating expenditure or investment, point to the activities that make the data and not the necessarily the entities that they set out to observe. Ultimately, these algorithmic configurations produce a 'mathematical formula that generates a number that can be compared to other numbers. It is singular and comparable, and consequently calculable, but in a way that is immediate' (Callon and Muniesa, 2003: 1236). These calculations, like,

> all forms of quantification (for example, probabilistic quantification, accounting quantification), transform the world, through their very existence, by their diffusion and use in argumentation, whether in science, politics, or journalism. Once the procedures of quantification have been coded and programmed, their results are reified. They tend to become "reality", by an irreversible "ratchet effect". The initial conventions are forgotten, the quantified object is naturalized, so to speak, and the use of the verb "measure" automatically springs to mind and into ink on the page. This naturalization remains in force until, for reasons that require case-by-case analysis, controversies erupt and the "black boxes" are reopened. (Desrosières, 2006: 7)

I suggest that large-scale digital assessments as tests such as PIAAC do more than produce data about ability. They perform the concept of ability into being. Ability is reformulated as a 'skill' (O'Keeffe, 2016) that can be framed as a calculable goods and ranked, valued and made transportable, as a type of 'resource that must be harnessed to underpin profitability' (Ozga et al., 2006: 6). Once set in motion as part of a test, the digital data produced by PIAAC on educational attainment are no longer 'merely an abstraction and representative, they are constitutive, and their generation, analysis and interpretation has consequences' (Kitchin, 2014: 21). Tests, like other 'scoring systems have the potential to take a life of their own, contributing to or creating the situation they claim merely to predict' (Keats Citron and Pasquale, 2014: 18) and their data production practices are all too easily black-boxed and rendered only partially accountable.

PIAAC was the first ILSA to be digital, but it will not be the last. Digital tests offer the possibility of large-scale, rapid data production of large amounts of data quickly and over time, much more cheaply than paper and pen assessments. Algorithms are depicted as neutral, efficient, objective, and trustworthy (Beer, 2017) and the data they produced as correspondingly value-free or objective. Efficient as they may be, test data are value laden and intricately meshed into the business and political goals of the organizations that deploy them (Gorur, 2014; Ozga, 2012). They play a role in creating a world where adults can be grouped according to ability and decisions about their lives made accordingly. Classifying people according to test scores valorizes specific kinds of knowledge that are easily testable and quantifiable (Bowker and Star, 1999). As such, assessment procedures have to power to 'legitimate certain types of behaviour in the education system—certain forms of knowledge, certain skills, certain attributes—as of value educationally and, by the same token, to devalue other kinds of behaviour as worthless or even undesirable' (Broadfoot, 2011: 121).

The emphasis on mass testing and scalability typical of large-scale assessment surveys has a decisive influence on educational and labour market policy and public enactments of accountability that value the predictive capabilities of big data produced by psychometric assessment. ILSAs such as those run by the EU, OECD, and IEA are 'premised on the presupposition that educational performance, like economic performance, can be monitored in terms of statistical fluctuations, risks, and country comparisons' (Williamson, 2015: 136).

With the development and naturalization of sophisticated and scalable digital testing assemblages, the emphasis on educational data shifts inexorably to skill measurement and algorithms play an increasingly central role in the 'deployment or expression of power' (Beer, 2017: 3) since the control of data production brings economic and political influence.

The data-producing practices of assessment test-makers are able to provide states and companies with data that reframes discussions about learning and education. Although the 'seemingly mundane, and often obscure, practices of collecting, storing, exchanging, processing, pseudonymizing or anonymizing digital data' (Bellanova, 2016: 3) often passes behind the scenes, these data and the technologies used to produce them have a profound and lasting influence not only on how adult education is performed and imagined but also on the lives of the millions who are categorized by assessment.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. In IRT, the skill vector is the sequence of items correctly or incorrectly answered and measures proficiency on a single skill. In this example the skills are literacy or numeracy. The latent cognitive profile is the sum of the correct or incorrect responses along the vector.
2. PIAAC, dependent upon as it is on IRT, uses linking items to calibrate the difficulty of each item. Using an Expectation-Maximization algorithm the test makers drew upon assessment data derived from tests administered in previous years such as IALS, ALL, or the PIAAC Field Test (Chen et al., 2016).

References

Appadurai A (1994) Commodities and the politics of value. In: Pearce SM (ed) *Interpreting Objects and Collections*. London: Routledge, 76–91

Baker FB (2001) *The Basics of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation

Ball SJ (2007) *Education Plc: Understanding Private Sector Participation in Public Sector Education*. Abingdon: Routledge.

Ball SJ and Junemann C (2012) *Networks, New Governance and Education*. Bristol: Policy Press.

Barber M (2004) High expectations and standards for all, no matter what. Creating a world class education service in England. In: Fielding M (ed) *Taking Education Really Seriously: Four Years Hard Labour*. London: Routledge, 17–42

Beer D (2017) The social power of algorithms. *Communication and Society* 20(1): 1–13.

Bellanova R (2016) Digital, politics, and algorithms: Governing digital data through the lens of data protection. *European Journal of Social Theory* 20(3): 1–19

Bondo Hansen K (2015) Review essay: The politics of algorithmic finance. *Contexto Internacional* 37(3): 1081–1095.

Bowker GC and Star SL (1999) *Sorting Things Out. Classification and Its Consequences*. Cambridge: MIT Press.

Broadfoot P (2011) *Assessment, Schools and Society*. Vol. 35. Oxon: Routledge

Brun-Cottan F, Forbes K, Goodwin C, et al. (Writers) (1991) *The Workplace Project. Designing for Diversity and Change [Video Tape]*. Palo Alto, CA: Xerox Palo Alto Research Center.

Callon M (2005) Peripheral vision: Economic markets as calculative collective devices. *Organization Studies* 26(8): 1229–1250.

Callon M and Muniesa F (2003) Les marchés économiques comme dispositifs collectifs de calcul. *Réseaux* 21(122): 189–233.

Callon M and Muniesa F (2005) Peripheral vision: Economic markets as calculative collective devices. *Organization Studies* 26(8): 1229–1250.

Carlson JE and von Davier M (2013) *Item Response Theory. ETS R&D Scientific and Policy Contributions Series*. Princeton, NJ: Educational Testing Service.

Chen H, Yamamoto K and von Davier M (2016) Controlling multistage testing exposure rates in international large-scale assessments. In: Yan D, von Davier A and Lewis C (eds) *Computerized Multistage Testing. Theory and Applications*. London: CRC Press, 391–410

Dean A and Kretschmer S (2007) Can ideas be capital? Factors of production in the postindustrial economy: A review and critique. *The Academy of Management Review* 32(2): 573–594.

Desrosiéres A (1993) *The Politics of Large Numbers. A History of Statistical Reasoning*. Cambridge, MA: Harvard University Press.

Desrosières A (2006) From cournot to public policy evaluation: Paradoxes and controversies involving quantification. *Prisme* 7: 1–42.

Domingos P (2015) *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake our World*. New York: Penguin Books

Dourish P (2016) Algorithms and their others: Algorithmic culture in context. *Big Data and Society* 3(2): 1–11

Fourcade M and Healy K (2016) Seeing like a market. *Socio-economic Review* 15(1): 9–29.

Geiger RS (2014) Bots, bespoke, code and the materiality of software platforms. *Information, Communication and Society* 17(3): 324–356.

Goldstein H and Wood R (1999) Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology* 42(2): 139–167.

Goody J (1975) *Literacy in Traditional Societies*. Chicago, IL: Cambridge University Press.

Gorur R (2014) Towards a sociology of measurement in education policy. *European Educational Research Journal* 13(1): 58.

Gorur R and Koyama JP (2013) The struggle to technicise in education policy. *The Australian Educational Researcher* 40(5): 633–648.

Graff HJ (2012) *Literacy, Myths, Legacies & Legends*. New Brunswick: Transaction Press.

Grek S (2014) Transnational education policy-making: International assessments and the formation of a new institutional order. In: Hamilton M, Maddox B and Addey C (eds) *Literacy as Numbers. Researching the Politics and Practices of International Literacy Assessment*. Cambridge: Cambridge University Press, pp.35–52.

Hamilton M (2012) *Literacy and the Politics of Representation*. Oxon: Routledge.

Hanushek E, Schwerdt G, Wiederhold S, et al. (2013) *Return to Skills Around the World: Evidence from PIAAC*. Paris: OECD.

Hogan A, Lingard B and Sellar S (2015a) Edu-businesses and education policy: The case of Pearson. *Professional Voice* 10(2): 24–29.

Hogan A, Sellar S and Lingard B (2015b) Commercialising comparison: Pearson puts the TLC in soft capitalism. *Journal of Education Policy* 31(3): 243–258.

Hogan J, Thornton N, Diaz-Hoffmann L, et al. (2016) *U.S. Program for the International Assessment of Adult Competencies (PIAAC) 2012/2014: Main Study and National Supplement Technical Report*. Washington, DC: U.S. Department of Education.

Hutchins E (1995) *Cognition in the Wild*. Chicago, IL: MIT press.

Keats Citron D and Pasquale F (2014) The scored society: Due process for automated predictions. *Washington Law Review* 89(1): 1–33.

Kitchin R (2014) *The Data Revolution*. London: Sage.

Lambert R and Leder S (2016) Les enseignants aux bons soins du patronat. *Le Monde diplomatique*. Available at: http://www.monde-diplomatique.fr/2016/11/LAMBERT/56790 (accessed 12 November 2016).

Latour B (2013) *An Inquiry into Modes of Existence. An Anthropology of the Moderns*. London: Harvard University Press.

Lingard B and Grek S (2005) The OECD, indicators and PISA: An exploration of events and theoretical perspectives. *Fab-Q working paper 2 CES*. University of Edinburgh, http://www.ces.ed.ac.uk/PDF%20Files/FabQ_WP2.pdf (accessed 23 May 2015).

Lingard B and Sellar S (2013) Globalization, edu-business and network governance: The policy sociology of Stephen J. Ball and rethinking education policy analysis. *London Review of Education* 11(3): 265–280.

Luke A (1997) New narratives of human capital: Recent redirections in Australian educational policy. *Australian. The Australian Educational Researcher* 24(2): 1–21

Mager A (2016) Search engine imaginary: Visions and values in the co-production of search technology and Europe. *Social Studies of Science* 47(2): 240–262.

McNamara T and Knoch U (2012) The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing* 29(4): 555–576.

Morgan C (2016) Testing students under cognitive capitalism: Knowledge production of twenty-first century skills. *Journal of Education Policy* 31(6): 805–818.

Neyland D (2015) Bearing account-able witness to the ethical algorithmic system. *Science, Technology & Human Values* 41(1): 50–76.

OECD (2011) *PIAAC Conceptual Framework of the Background Questionnaire Main Survey*. OECD: Paris

OECD (2013a) *OECD Skills Outlook 2013: First Results from the Survey of Adult Skills*. Paris: OECD.

OECD (2013b) *Technical Report of the Survey of Adult Skills (PIAAC)*. Paris: OECD.

O'Keeffe C (2015) *Assembling the adult learner: International and local assessment practices*. PhD Thesis, Lancaster University, Lancaster.

O'Keeffe C (2016) Producing data through e-assessment: A trace ethnographic investigation into e-assessment events. *European Educational Research Journal* 15(1): 99–116.

Ozga J (2012) Governing knowledge: Data, inspection and education policy in Europe. *Globalisation, Societies and Education* 10(4): 439–455.

Ozga J, Seddon T and Popkewitz TS (2006) *World Yearbook of Education 2006: Education, Research and Policy: Steering the Knowledge-Based Economy*. London: Routledge.

Pasquale F (2015) *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard: Harvard University Press.

Piketty T (2014) *Capital in the Twenty-First Century*. Cambridge, MA: Belknap Press.

Rieder B (2016) Scrutinizing an algorithmic technique the Bayes classifier as interested reading of reality. *Information, Communication and Society* 20(1): 100–117.

Rose N (1996) *Inventing Our Selves: Psychology, Power and Personhood*. New York: Cambridge University Press.

Slater D (2002) Markets, materiality and the 'new economy'. In: Metcalfe S and Warde A (eds) *Market Relations and the Competitive Process,* Manchester University Press: Manchester, 95–113.

Thorn W (2014) "Riding on unfinished rails" – designing PIAAC. LLinE: Lifelong Learning in Europe 1:1–9. Available at: http://www.elmmagazine.eu/articles/riding-on-unfinished-rails-designing-piaac (accessed: 4 December 2014).

Verger A, Lubienski C and Steiner-Khamsi G (2016) *World Yearbook of Education 2016: The Global Education Industry*. London: Routledge, 219–228.

von Davier M and Ying C (2016) Multistage testing using diagnostic models. In: Yan D, von Davier A and Lewis C (eds) *Computerized Multistage Testing. Theory and Applications*. London: CRC Press

Wang H, Cuiqin M and Chen N (2010) A brief review on Item Response Theory models-based parameter estimation methods. In: *Paper presented at the 5th International Conference on Computer Science & Education*, (19–22). IEEE: Hefei, 24–27 August 2010

Williamson B (2015) Digital education governance: Data visualization, predictive analytics, and 'real-time' policy instruments. *Journal of Education Policy* 31(2): 123–141.

Williamson B (2016) Boundary brokers: Mobile policy networks, database pedagogies, and algorithmic governance in education. In: Ryberg T, Sinclair C, Bayne S, et al. (eds) *Research, Boundaries, and Policy in Networked Learning*. Springer

Yamamoto, K. (2011, September). Implementation of CBT in the PIAAC field test and CAT in the PIAAC. In: *20th seminar report at the Centre for Research on Educational Testing*, Tokyo.

Zabal A, Martin S, Massing N, et al. (2014) *PIAAC Germany 2012: Technical Report*. Münster/New York: The Federal Ministry of Education and Research.

Ziewitz M (2016) Governing algorithms: Myth, mess, and methods. *Science, Technology and Human Values* 41(1): 1–14.

Author Biography

Cormac O'Keeffe earned his Ph.D degree in Education from Lancaster University's Centre for Technology Enhanced Learning. He is currently the executive director of training and assessment at YES 'N' YOU in Paris. His research interests are STS, digital education, learning analytics, and software studies.