Technologies and infrastructure: costs and obstacles in developing large-scale computer–based testing

Gavin T. L. Brown

To cite this article: Gavin T. L. Brown (2018): Technologies and infrastructure: costs and obstacles in developing large-scale computer–based testing, Education Inquiry, DOI: 10.1080/20004508.2018.1529528

To link to this article: https://doi.org/10.1080/20004508.2018.1529528

© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 16 Oct 2018.

Submit your article to this journal

View Crossmark data
ABSTRACT
This paper identifies important technological, financial, and equity challenges to the development and implementation of online or onscreen educational testing. The insights are based on the author’s experiences in New Zealand of being the senior project manager for the Assessment Tools for Teaching and Learning (asTTle) computer-based standardised testing system. The system was funded by the government for formative educational purposes and is delivered to schools for free. The issues faced in the project that need to be considered in any new online testing system include: clarity about the purpose of the testing system, the importance of knowing the current technology conditions, the challenges of working within the dynamic and changing world of technology, the need to assure quality and equity for all test-takers, the economics of paying for a public, rather than commercial, service, and challenges in the management processes of developing and introducing a new system. This survey of technology issues makes it clear that moving to online or onscreen testing has many fishhooks in it that need substantial thought prior to engaging in the development process.

KEYWORDS
Computerised testing; agile development; equity; quality

The conversion of paper-based standardised tests to computers or even online poses some interesting and challenging difficulties. While it may be considered rather straightforward to display a paper test on a screen and capture the options test-takers select in a multiple choice format, there are some challenges that are frequently overlooked and which may prevent the grander dreams of computer-enabled testing from taking place. The purpose of this paper is to outline some of the issues by illustrating experiences and lessons learned from New Zealand’s e-asTTle test system.

Testing in New Zealand
New Zealand schools have a long tradition of using standardised testing for diagnostic and formative purposes (Crooks, 2010). For example, the Progressive Achievement Tests1 were developed in the 1960s by the New Zealand Council for Educational Research to provide teachers with a way to annually monitor student achievement in
a range of key content areas (i.e., reading comprehension, reading vocabulary, mathematics, listening skills) calibrated against national norms. The use of these tests was determined by each school, the data were retained in the school, no national reporting was required, and the information was meant to guide instructional programming for students and inform reporting to families. Over the following decades, a number of diagnostic and formative assessment tools have been added to the assessment tool-box (see details in Brown, Irving, & Keegan, 2014), including the Supplementary Test of Achievement in Reading, the BURT Word Reading Test, and the School Entry Assessment Kit. All of these tests are low-stakes, school-controlled, and designed for formative use by teachers, especially for guiding instructional planning and discussing student learning needs with parents and families.

In the 1990s, the New Zealand Ministry of Education introduced an eight-level curriculum framework using an outcomes-based approach that identified essential learning areas, strands of learning objectives, and levels of progress (Ministry of Education, 1993). This was generally welcomed by classroom teachers (Duthie, 1994; Lennox, 1996). At the same time, the Ministry began initiatives to improve the capacity of teachers to assess student learning against the objectives of the curriculum. These resources included additional assessment tools (e.g., Assessment Resource Banks for reading, mathematics, and science; Croft, 1999) and teacher professional development services (e.g., Assessment for Better Learning; Education Gazette, 2002).

Consistent with long-standing practices, teachers were expected to report student progress within this curriculum framework to families (Ministry of Education, 1994). Furthermore, within this educational framework, there were simultaneous calls for greater regulatory control expressed through government-mandated uses of assessment. These included the National Educational Goals, National Administration Guidelines, the Education Standards Act of 2001, and the National Standards 2010 (rescinded in 2017) legislation. These mechanisms sought to establish commonality across the approximately 2400 schools of New Zealand and greater efficiency and effectiveness in achieving government goals for all children. While the current Minister of Education seems to be opposed to measurement and accountability, this does not necessarily translate into whole-scale rejection of tests which inform teachers and parents about who needs to be taught what next (Brown & Hattie, 2012). Indeed, it is clear that New Zealand teachers do not have antagonistic views towards tests in and of themselves, if they perceive that test information contributes to improved teaching and student learning outcomes (Brown, 2011). Thus, a whole-scale rejection of tests as part of the repertoire of educational assessment practices within New Zealand is unlikely.

To help schools meet these targets, the Ministry of Education funded in 2000 the development of a new externally referenced (norms for Years 5–7; nominally ages 10–12) testing system of literacy and numeracy in both English and Māori languages that was simultaneously indexed to curriculum levels 2–4 (Years 4–8). A call for proposals resulted in the design from the University of Auckland, led by Professor John Hattie, being selected as the preferred provider, with first work being done in August 2000. The asTTle system was designed as a standardised multiple-choice test bank that would be controlled at the school-level to (a) assist in better diagnosis of student learning needs, (b) inform improved curricular instruction in the classroom, (c)
provide for more informative reporting to parents, and (d) generate more accurate plans and targets for school-wide improvement.

The emphasis on externally referenced or norm-referenced information may seem out of place within a formative assessment agenda. However, because schools are situated within communities, it is difficult for teachers to know where their own students sit relative to the performance of students across the whole nation. Normalising on one’s own experience often means that teachers in high socio-economic communities are prone to thinking their relatively weaker students are actually weak, leading to potentially inappropriate reports and actions. This may be a representative heuristic at work (Kahneman, 2011) that can negatively impact teacher judgements about students based on their relative position within a group, rather than on the relative position of the group vis a vis the nation. Hence, having access to what is average across student demographic characteristics (i.e., school year, sex, ethnicity) and school characteristics was deemed useful.

The New Zealand e-asTTle system

It is worth noting, then, that the applicability of the issues outlined in this paper to other development initiatives will depend in part upon the intended goals of the test system and its funding. The New Zealand system was developed, under government contract, as a diagnostic and formative testing system to be used in schools without reference to an external testing agency. This is in contrast to systems that evaluate schools and teachers by state-mandated accountability testing (Darling-Hammond, 2003) or which administer tests at set-dates by external agencies to monitor the performance of a system (e.g., the National Assessment of Educational Progress in the United States) or to make international comparisons (e.g., the Programme in International Student Assessment). Even more, it stands in contrast to certification tests by which individuals are tested for entry to higher education (e.g., Educational Testing Service’s GRE) or into a professional status (e.g., the Architect Registration Examination). The New Zealand system was fully funded by the government and is still provided to schools free. This contrasts to testing programs that are funded by private corporations or trusts or through direct charging of test-takers. These factors change some of the dynamics about to be described.

e-asTTle is a nation-wide, internet-delivered standardised test system for reading comprehension, mathematics, and writing. The system has been described in a number of peer-reviewed publications (Archer & Brown, 2013; Brown, 2013; Brown, O’Leary, & Hattie, 2018; Hattie & Brown, 2008; Hattie, Brown, & Keegan, 2003; Hattie, Brown, Ward, Irving, & Keegan, 2006; Keegan, Brown, & Hattie, 2013). Hence, this summary provides a high-level description of the system, with interested readers encouraged to review the listed publications.

e-asTTle is curriculum-aligned, covering achievement objectives for levels 2 to 6 (normally taught in Years 4 to 12, inclusive) (Ministry of Education, 2007). It is school-controlled; teachers and school leaders decide whether, when, and how the tests are designed, administered, and interpreted. The system contains a menu of test design options (e.g., length, administration mode, fixed or adaptive, selection of attitudinal question topics) selected according to the analytic need of the test administrator, who is
frequently a classroom teacher. Use of e-asTTle is voluntary, with about half of all schools in New Zealand currently using e-asTTle on a regular basis. e-asTTle is diagnostic of student strengths and weaknesses within each test, and provides additional information of performance relative to national or local regional normative performance statistics. This is done through graphical reports (Brown et al., 2018) at the individual level or for complete classes or school cohorts according to the needs and interest of the test administrator. Diagnostic capability is achieved through item-response theory pre-calibration of an extensive bank of test items and the determination of student scores using a one-parameter logistic statistical model.

Reports of performance and need are generated automatically when students complete the test and a menu of test report options (e.g., individual vs. group; norm-referenced vs. criterion referenced) is available according to the analytic need of the test interpreter. Computer adaptive adjustment of test difficulty is an option for reading comprehension and mathematics, though test-designers can select fixed form delivery on either paper or screen. The system is designed to be formative in that it connects student learning needs to an online catalogue of teaching resources (called What Next?\textsuperscript{16}) that are aligned to the curriculum achievement objectives and levels covered by the tests themselves. This means that test results can be responded to with additional resources aligned to detected needs.

The tests are custom-designed by each teacher or school leader by choosing test length, achievement objectives, and curriculum levels. An underlying simultaneous linear program maximises the choices of the test-creator and minimises undesirable test characteristics (Fletcher, 2000), such as previously used items from the test bank or the absence of cognitively deep item content. The Ministry of Education provides web-hosting and internet access for all schools, so that there are technically no obstacles to using the system at any time during the school day, week, or year. The e-asTTle system was developed gradually and incrementally from 2000 to 2007 with extensive interactive and formative evaluations (Hattie et al., 2006) when the current e-version was deployed. That system has now been running continuously for a decade. Hence, the lessons learned in developing the technology and infrastructure for computer-based large-scale testing can be informative for other potential systems.

**Technology infrastructure**

A key constraint of on the design of the asTTle system was the funder’s requirement that it function within the technology environments of schools. In the interests of equity, it had to run on the lowest common denominator of hardware and software available in schools. Unfortunately, the Ministry of Education’s understanding of the machine infrastructure of schools was not comprehensive or accurate. This arose largely because the purchase of computers and their installation within schools was the responsibility of schools themselves, making use of operational funds for furniture and equipment or from parental donations or fund-raising (Fiske & Ladd, 2000). Schools are expected to raise funds from their parent communities or philanthropic sources to provide high-quality computing for students and teachers. Consequently, the quality of information technology systems available in schools lagged considerably behind that available in industry or commercial enterprises. I suspect that few
jurisdictions have better IT in schools than in business; hence, this will provide a general constraint on all online testing.

At the time asTTle was being formulated (2000–2002), the government did not have minimum standards, reporting mechanisms, or funding by which they could control the type of computing environments in schools. This meant that there was great diversity in types of machines, their age, their operating systems, and their local area network design. In the early 2000s, schools were still running quite old computers with limited capabilities. For example, asTTle versions 1 and 2 were required to operate on Windows 95 and Mac OS 8 operating systems. Also, because the government had not deployed high-speed broadband internet to all schools across the country, the system had to run on standalone machines, and by asTTle version 4 on local area networks within schools.

Hence, the system had to run on machines with limited graphic or processing capabilities. This means that the visual appearance of the system had to have low demand on memory and processing power, resulting in quite simple colour palettes and few on-screen animations. Indeed, in asTTle version 4 a rotating icon was introduced to reassure users that the system was working to create a test or process data (Hattie et al., 2004). As the system was deployed into secondary schools, however, support for the older machines was abandoned, as it was found that those machines could not cope with the ever-decreasing processing times expected or tolerated by users. This suggests, that aiming to support the oldest systems is a politically desirable ambition, but may well be economically unjustifiable.

Another infrastructure barrier had to be addressed in versions 3 and 4. School leaders and middle managers wanted to be able to see what the testing results were from various classes that had administered different asTTle tests. The asTTle system allows comparison of performance even when students do different test forms because all items in the bank are calibrated to a common underlying ability scale with item response theory statistics (Wright & Bell, 1984). Thus, ability estimations can be compared despite students completing different test forms. If data were resident only on individual teachers’ machines, including possibly a laptop, it would not be possible for the manager to get an overview of achievement or needs across a range of classes or cohorts.

Hence, the Ministry supported the development of local area network versions of the test system so that all student scores and all test forms could be seen by any legitimate user of the network. However, again variations in the school environments meant that the original contract to build network versions for Windows, Mac, and Linux servers was insufficient. A small number of secondary schools had privately developed a consortium of Novell network users and a new contract had to be let to design and test that the asTTle system worked on Novell networks. It became evident very quickly through this five-year process that creating a system that depended on the multiplicity of school infrastructures was no longer feasible. Instead, a two-year project was initiated to convert the asTTle LAN system to a completely online internet hosted e-asTTle.

Under current technology infrastructure systems, this approach could be completely by-passed by moving straight to a cloud-based approach. The major advantage of a hosted Internet service is that all development costs are standardized around Internet browsers and a single hosted server that delivers tests and stores data on the half of school-based users (Armbrust et al., 2010). This structure also reduces costs for the
client as processing power is held at the cloud-based server. No longer does it matter what type of computer operating system or local network the school has because all users can log on to the Internet and make use of the hosted service. Nonetheless, while such a solution seems obvious there are hidden traps to do with bandwidth, server processing, and security.

The asTTle development team recommended the use of an external professional hosting company for a number of reasons (i.e., relatively low start-up cost, high levels of data and access security, robust ability to cope with service demand, and visible independence from the Ministry of Education). However, the Ministry of Education decided to create its own dedicated service. Unsurprisingly, given that IT infrastructure is not the core business of the government’s policy body in education, there was considerable delay (approximately one year) before the online system could go live.

Furthermore, perhaps due to budget limitations or mis-estimation of demand, even now there are time points in which administration of student testing or the processing of report analysis is considerably delayed. Imagine that 50% of New Zealand’s 2400 schools in the country want to administer a test to approximately 30 students (i.e., one class) at the same time (a scenario that may occur at the start or end of school year or school term), this would result in 1200 * 30 = 36,000 concurrent users. Users receive error messages when the system is overloaded, meaning online tests cannot be administered. This has, depending on the time of school year, impacted use of e-asTTle in a current research project dealing with 9–10 year olds born hypoglaecimic. This gives developers a need to consider what realistic peak demand might look like.

An additional benefit of moving to an online system is the removal of hand-marking by teachers. However, because not every school necessarily had sufficient computers to be able to test as many students online as they wanted at any one time, the option of paper-based testing had to be maintained throughout the design and deployment of the system. Even the current e-asTTle has the option of paper-based testing, despite the additional workload of hand-scoring and data entry. Ensuring technology environments in which tests are deployed exist so that the features can be exploited is a responsibility which is often devolved and underfunded.

Consequently, given the age of the New Zealand asTTle technology (i.e., a decade old) and the infrastructure challenges apparent in the system, the Ministry of Education has been seeking technology investors to improve not only the end-user experience, but also the complete technology infrastructure so as to achieve expected end-user experience. Overlooking the legitimate expectation of end-users about no speed or time barriers to going online and administering or processing tests is a serious challenge for the development of large-scale online testing. Without building for peak rather than average use, an online system is likely to be perceived negatively and underutilized.

It is worth considering that these system constraints may no longer apply in societies that have moved rapidly to high-speed broadband, full mobility, and nearly universal access to touch devices. Nonetheless, in moving to online testing, awareness of these constraints is necessary to ensure appropriate design. We have recommended elsewhere (Hattie & Brown, 2008) that a gradual incremental approach should be taken. However, this does not mean new systems have to start with stand-alone or local area network versions; new systems need to begin with relevant current technology environments. However, it still seems logical that, as with any fundamentally new system,
developments will be incremental and funding will be required to update original designs in keeping with technology changes.

**Changing technology**

One of the truisms of computer technology is that it is constantly changing (Schaller, 1997). This had significant impact on the design, development, and current status of e-asTTle. During the development from asTTle v2 to asTTle v4, there was approximately two years between final testing and submission to the client of each version of the fully working software. During those two years, Apple changed its Mac OS operating system 23 times. In two of those changes, the software protocol for how third party applications launched PDF documents changed. This had a significant impact on asTTle because both the tests and score reports were displayed as PDF documents. In both cases, by the time the client’s own technical staff tested the software (some 3–4 weeks later), they found that the Mac version did not work and developed a very negative view of the software. While these problems were not very difficult to fix, the changes necessitated a complete suite of testing on all machines, operating systems, and networks to ensure that in fixing this one problem, no new problems had been introduced. This tiny example illustrates a much larger problem.

In the time since asTTle v4 went live (January 2005), the software and hardware environment of the world have changed immensely. Touch screens, mobile devices, virtual reality, social networks, gamification, and so on have all been introduced and to some degree become commonplace. Current developments in on-screen testing are taking advantage of natural language processing in which students can type responses to an artificially intelligent computer avatar that interacts with the test taker (Katz & Gorin, 2016). As technology develops, becomes more ubiquitous, and less expensive, end-users will have rising expectations for what online testing should be able to do.

A challenge to integrating advanced features into testing software resides in the expectation that all children or adolescents have equivalent opportunity to learn or experience the types of software being utilised to test learning. Recent US studies have reported that having a home computer and internet access for longer (i.e., from before start of schooling) is associated positively with achievement (Robinson, Wiborg, & Schulz, 2018; Vigdor, Ladd, & Martinez, 2014). A pre-schooler in a relatively wealthier home in contemporary western societies is likely to be exposed to home computers and the internet. Nevertheless, young children, even in an urban, low-income, minority American community, have almost universal exposure to mobile devices (including touch-screen devices, such as smart phones or tablets) and most had their own device by age four (Kabali et al., 2015). This suggests that contemporary students may have reasonably robust familiarity with information technologies, but it does not necessarily imply they are familiar with the tools available at school or in an online testing system.

If a testing system requires extensive use of a pointing device such as a pen or mouse, instead of a touch screen, then the system has to ensure all children know how to use the pointer and have the fine motor control skills to manipulate such tools. Alternatively, touch screen methods for indicating multiple-choice question option selection may be feasible, but will still require some careful calibration depending on screen size and number of items displayed on a screen. Fundamentally, not everyone
has the same degree of finesse in touching or manipulating objects by touch on a screen
and not every touch device is equally sensitive or responsive.

In a second example, consider the development of innovative item types. The asTTle
system began with a very ‘vanilla’ display of items and the presentation of item
responses (i.e., a MCQ radio button had to be clicked to indicate selection of an
option). During asTTle v4 development, the software team demonstrated, for the
purposes of assessing learning in mathematical measurement, the ability to create
transparent rulers and protractors that could be dragged, dropped, or rotated to
simulate the real world exercise of measuring and recording the length or characteristics
of shapes. This process required very fine motor control skills to manipulate the
‘handles’ on the measurement tool to position it and read off the correct answer. The
ability to zoom in on the object would make such manipulation easier, but also requires
the user to know how or when to use a zoom device (conventionally signalled with a
magnifying glass icon, even though increasingly fewer people know what a magnifying
glass actually is). The whole process was shelved, despite being technically feasible more
than a decade ago, for the amount of time it would take to prepare each item versus the
possibility of preparing many more ‘vanilla’ items in the same time.

This raises the trade-off question about the relative virtue of realism or authenticity
in a computer administered test versus obtaining larger or more representative samples
of student knowledge through less-authentic test items. The use of novel technology
features seems attractive because of the potential to more fully represent a construct and
a test-taker’s abilities within that domain (Sireci & Zenisky, 2006). Modern technology
allows the test-taker to interact with an on-screen teacher avatar who can respond to
the natural language of the student (Katz & Gorin, 2016). Such features certainly mimic
the reality of a classroom setting and may permit students to show their real abilities.
However, if the cost and time needed to implement and train test-administrators and
test-takers in using novel assessments is substantial, it may be more prudent to use the
same resources to develop many more items that ensure better coverage of the domain.
Online test developers in any new setting will need to weigh up the competing
challenges of what technology can do with what accurate score report generation
might require.

**Economics of testing**

Significant costs in building modern testing software lie in the development of test
content and in the norming of items. Rudner (2009) has estimated the conventional
development cost per item for a large-scale standardised test to be between US $1500
and $2500. This pays for item writing and reviewing, copyright fees, graphic design,
preparation of test booklets, field administration, statistical analyses, and standard
setting. Thus, overcoming the cost of new items is a significant objective, so that
more can be spent on infrastructure and professional development.

In moving to computer assisted, online testing, many organisations may already have a
pool of pre-calibrated items to integrate into a new system; hence, initial item development
costs may be quite low. Further, new methods of automated item generation might drastically
reduce the cost of item development (e.g., Gierl & Lai, 2016, 2018). Automatic item genera-
tion is especially important when new items are needed for each round of an annual test for
school accountability or quarterly administration of professional certification examination. Of course, automated item generation techniques are not a solution for all subject areas or for all types of learning outcomes (e.g., performance on complex skills). In contrast, in jurisdictions where online testing is being developed for formative in-school use (as the e-asTTle system was), end-users may have less confidence in items which were not developed and evaluated by expert teachers. Indeed, it was a factor in the successful adoption of the e-asTTle system that a list of teachers who had participated in item writing, reviewing, calibration, or standard setting was included with the manual (Hattie et al., 2004). Teachers could see the names of others whom they knew and respected and this helped generate acceptability. Consumers are used to paying less and less for new information technology devices and receiving many useful software systems at no cost. However, this does not mean that those technologies cost nothing to develop. Indeed, technology costs are not zero, nor necessarily trivial. There are substantial costs in developing a software system that administers and scores items, generates reports, stores and links data for longitudinal profiling of test taker proficiency across time and subjects, communicates efficiently across the internet upon demand, and which keeps all student, teacher, and school data secure. Unfortunately, technology development projects involving new software are as likely to fail as succeed or experience significant cost and time over-runs (Nasir & Sahibuddin, 2011). The world of continuously changing information system software and hardware options, features, and capabilities threaten the viability and feasibility of many projects, let alone any changing scope and goal decisions made by project funders and/or sponsors. Thus, strong management systems are needed to properly estimate costs, manage expectations, and ensure delivery meets base requirements. All of these risks mean that fixed price estimation is difficult and often contains a high degree of risk.

In light of the recent Facebook–Cambridge Analytica personal data scandal, it is extremely important that student test data are kept secure (Tierney & Koch, 2016). These are all features that banks, shopping websites, and travel companies deliver. These costs do not have much impact on video and computer gaming companies or financial institutions because they are able to attract substantial investment to make their software efficient, effective, and economical for the user. Ultimately, they sell their products directly to the consumer or through advertising (e.g., Facebook). However, government agencies that want online testing are much less likely to be in a position to charge end-users for the very tests that they require children or young adults to take. If taking the government test requires paying a fee for an online administration and there is no subsidy, then large inequities are introduced by law or regulation. While ETS can safely charge its customers for their certification tests (e.g., GRE or TOEFL), the same cannot be said for Grade 6 national mathematics or reading tests. Thus, if online testing is to be part of the compulsory educational system, substantial seed and continuing funding is going to be needed.

Development is not the only cost. Maintenance is an ongoing cost. Unlike the development of a new curriculum-aligned text book, there are continuing costs for providers of testing software. Once a textbook or teacher professional development program is designed, authored, printed, and published, the only ongoing costs are reprinting and dissemination. In contrast, software requires constant updating as the technology infrastructure evolves, as described earlier. It is illustrative that in New Zealand, the asTTle test system was initially housed in the Curriculum Division. This
was a logical place since the tests were designed to measure progress within the curriculum. However, the Curriculum Division had no expectation or experience that they would have to fund support over the long-term for a team that could fix the software if something changed in the technology environment. Indeed, as the system was being deployed across all schools, it was only by invoking the metaphor of fire insurance (i.e., something you pay for hoping that nothing goes wrong, but purchase just in case it might) that persuaded the Ministry officials that they had to fund a technology team through the initial deployment processes.

A final cost, often overlooked by developers of cloud software, is the last few inches in the technology chain. The end-user experience of an online test depends on the terminal connected to the internet and its display interface. A review carried out as part of the transition to online testing made it clear that what made online testing equivalent to paper-based testing was the quality of the visual display (Leeson, 2006). When students are presented with paper tests, they can easily see two pages simultaneously so that questions can be physically placed opposite or near stimulus material (e.g., reading passages). Without significant strain, in a paper format, the reader can zoom and scroll between questions and items and see the printed diagrams and text in high quality display. In contrast, computer monitors are normally smaller than an A4 or 8½ x 11” sheet size. This means that the reader cannot see the whole page and all the test questions and reading passage in one glance. Leeson’s (2006) review identified that the extra labour of scrolling or zooming to find answers in stimulus material and switching back to test items created an additional functional burden on test-takers. Although computer screens and software have zoom and scroll features, how they do this varies between manufacturers and software designers, creating extra burdens for ensuring equitable familiarity with the test infrastructure.

Further, computer screens tend to be designed in landscape format rather than portrait which is the general default for displaying tests. This means that test items might have to be displayed to the side of reading material. Also it has implications for input systems, especially keyboards, because rotating a tablet to portrait restricts the space available for the on-screen keyboard. Additional on-screen prompts (e.g., thumbnails of passages and items) might be needed to support the user’s transitions among items in a testlet and the related stimulus material.

In a paper-based test, users can flick between pages, return to items, and easily change their responses. Hence, provisions have to be made to allow test takers to similarly go forward and return, submit answers, and change them. The technology has to record and store the item selection, but then allow the test-taker to return to the item, review, and possibly change a previously given answer. This means that, in the spirit of equity of opportunity, test developers may have to establish and ensure standards for end-user machinery and visual display units. Without such assurance, perhaps the online or onscreen testing may not be perceived as valid.

An advantage of computer-testing is that the process information itself (i.e., navigation between items, changes in answers, time between items, etc.) all become potential information to discover not only what answer was given, but what skills and abilities were exercised in reaching the answer. This feature is already exploited in systems that monitor the response time students take to answer computer-test items and provide prompts to slow-down if rapid guessing is detected (Wise & Smith, 2016).
possibilities that process-data analysis give to testers is one of the key promises of online testing (von Davier, 2018).

This means the cost of placing computers in front of every test-taker has to be borne somewhere. In many societies (e.g., Sweden, New Zealand), parents or schools bear the costs of providing computing for students. For example, the ubiquitous presence of Google Chromebooks within bring your own device (BYOD; McLean, 2016) educational technology solutions means that children from reasonably well-off homes can afford devices, but this may not be a universal assumption. Lack of access to a machine by which a child can be said to be familiar with the online testing environment clearly introduces a construct-irrelevant factor in the test scores.

Furthermore, in the interest of making BYOD affordable, short-cuts might be taken in terms of visual display unit size or resolution, or processing power of the hardware (e.g., the quality of memory, processor speed, and visual display in lower cost machines, more likely to be purchased by or for school, tends to be limited)\(^1\). Hence, for equitable testing purposes, a testing agency may have to invalidate testing that does not take place on sufficiently high-quality devices. It may be difficult to persuade or ensure parents or schools to pay for this standard, especially if it were only necessary for the administration of government-mandated testing. In social democratic societies, citizens might have a legitimate expectation that government-mandated testing is carried out on government-supplied software and hardware. Any inequity in providing required technology solutions may create significant construct-irrelevance in the scores and substantial tensions among voters.

Thus, the assumption that online testing can efficiently and effectively piggy-back on the already substantial community investment in ICT, may actually invalidate not only the scores but the world of technology-enabled testing. Contrariwise, citizens exposed to advanced technologies may wonder why online tests are so ‘vanilla’ in lacking the advanced features they experience in their own every-day use of technology. Of course, asking citizens to pay more taxes or move funding in government budgets from health, for example, to testing technology may not be vote-winners either.

**Project management in technology**

The asTTle test system was structured as a research and development project. The development of test questions for specific achievement objectives and levels of the curriculum was new to the country, although standardised multiple-choice tests of generalised ability in reading and mathematics with reporting against norms have long been part of the New Zealand school system. The development of a computer-assisted test design, analysis, and reporting system was new. The government’s call for proposals did not have specifications or standards by which the technology should be designed. Hence, the project decided not to head in the direction of innovative item types, although this has been strongly recommended as an important direction in modern testing practice (Pellegrino, Chudowsky, Glaser, & National Research Council, 2001).

Given the novelty of introducing a computer-assisted mechanism for customised test creation (instead of fixed forms) and for reporting diagnostic information to teachers, instead of psychometricians, the whole project was designed as two interactive research
streams. One stream focused on an analysis of the curriculum to inform the design of item types, their difficulty, and scoring process. The goal of this stream was to ensure that the items had validity to both the curriculum and classroom contexts. The second stream was to develop a communication framework through which the technical knowledge of psychometric statistical analysis could be effectively communicated to teachers who had to use tests to identify which students needed to be taught which content at what difficulty next (Brown et al., 2018). Additionally, the system had to address the multiple objectives teachers and school leaders had for giving a test (i.e., guiding instruction for individuals and groups, advising curriculum managers and school leaders about the efficacy of teaching programs, informing students and families about individual learner’s strengths and needs, connecting tested performance to curriculum teaching resources, etc.).

Within these evolving research and development streams, the design and testing of an effective software system was conducted. Given that there were no fixed standards or exemplars on which to model the system, a narrative requirements specification was developed and used by a software development company to inform coding (Hattie, Brown, & Keegan, 2001). This meant that instead of providing a complete set of specifications to a coding house, the asTTle project team and the software company used an agile development process to interactively clarify the meaning of the requirements and to enact and test the resulting code (Dingsøyr, Nerur, Balijepally, & Moe, 2012; Highsmith & Cockburn, 2001). Agile development works well in a research environment because the exact details of what is needed and will be delivered remain somewhat fluid, while deadlines and budgets are preserved.

The approach taken with each version of the software followed a similar format. A list of possible features and desiderata were prioritised by the asTTle team and the Ministry officials. The implementation cost of each item on the list was priced and a decision about relative priority was made using the simple mnemonic MOSCOW (standing for MUST, SHOULD, COULD, and WOULD) (Hatton, 2007). Only the items listed as MUST were guaranteed for development by the due date for the quoted price. However, additional items from the SHOULD list would be delivered if estimates of effort on the MUST list had been overly pessimistic; whereas, none of the items classified as COULD or WOULD were ever going to appear without a significant increase in time and/or budget. This priority setting approach gives recognition to many great ideas that cannot or possibly should not be implemented, especially in early versions of the system, when acceptability has not yet been established.

Every three months (referred to as a ‘time-box’), progress on the agreed list was demonstrated, with opportunity for the asTTle team or the Ministry sponsors to revise or approve. The fewer revisions required, the more items on the SHOULD list could be obtained. The relative priority of or the introduction of new requirements was revisited at these quarterly review time-points. This approach gives considerable flexibility in the design and revision of the software design process, especially important as initial versions were evaluated in the ‘wild’. Real-life feedback from users can be greatly powerful in designing a new system. However, after three time-boxes, no new specifications or features could be added because the last quarter of each year’s objectives was to integrate all new features with new data and the previous version so that full testing could be accomplished.
What this agile, time-box, prioritised approach meant is that the exact list of features to be delivered could not be specified until after 9 months of software development. The minimum MUST list would be present as would a wide variety, but unspecified number, of SHOULD features. In the asTTle experience, this approach seemed to violate the conventions normally adopted by organisations such as IT development governance boards, which focus on control and mitigation of risks (Hardy, 2006). These bodies tend to insist on knowing in detail exactly what is going to be delivered for the purchase price. To utilise such a standardised process would, however, require the software company to wear all the risks, greatly inflating the price to cover the unknowns.

Given that in the first five years of the project, four fully working versions of the asTTle software were delivered on-time and on-budget stands testament to the superiority of this agile approach. Given that many software development projects fail to deliver as originally specified (Nasir & Sahibuddin, 2011), the accomplishments of the asTTle team in creating a fully functioning, teacher-controlled, school-based testing system is commendable. In an agile development framework, exactly what is needed is not known initially, so the temptation to constantly change or modify requirements is considerable. This approach controlled scope creep, which happens when sponsors or clients seek to add more requirements without a change in the time or budget framework.

Thus, considerable political work has to be done to ensure that a government funder can work with this agile approach. An agile approach is probably essential in devising an online testing system because each society has different values and attitudes towards testing and each government has different objectives in introducing computerised testing. A major contributor in the asTTle system has to be that it was not primarily an external accountability tool designed to punish teachers for students’ poor test scores. An agile approach allows interaction with intended users and actual deployment environments. Treating this as a flexible discovery-learning problem seems highly recommended. If we can learn one thing from the New Zealand experience, it is that it takes time to introduce a society to new methods of testing educational learning. Resistance to novelty is common-place (Struyven & Devesa, 2016) and an incremental process is likely to win the hearts and minds of intended users.

**Conclusion**

This short paper, taking advantage of insider knowledge from working for five years on the development of New Zealand’s computerised testing system, has identified a number of challenges in introducing large-scale computerised testing in any society that does not already have a history of such a process. The key ideas I have outlined are:

- Clarity about the purpose of the testing system,
- The importance of the current technology conditions,
- The dynamic and changing world of technology,
- The need to assure quality and equity,
- The economics of paying for a public, rather than commercial, service, and
- Challenges in the management of developing and introducing a new system.
The important questions these challenges raise have to do with ensuring equity of opportunity to learn how to interact with the system and ensuring quality of experience with the system so that construct-irrelevant factors do not contaminate resulting scores. Elsewhere, we have argued for an incremental approach (Hattie & Brown, 2008) in moving from a very simple implementation to ensure acceptance of the whole system, before moving into the innovations that modern technologies can offer.

The lessons learned from the asTTle project depend largely on the effort by the government to create a better kind of standardised testing system. Instead of tests that only reported total score and rank order, which had been the status quo for decades, the asTTle system attempted to link testing with teaching and curriculum by giving teachers a mechanism to design tests according to their own understanding of the curriculum and obtain new insights as to student learning needs in accordance with the very same curriculum framework that was guiding their instruction. Further, by automating analysis of achievement, the tests could point teachers to further curriculum resources and so further close the feedback loop from teach to test to teach via the curriculum.

Hence, because the goal of this online testing tool was to support learning and teaching, the lessons reported in this paper may have much less applicability to large scale testing programs which aims to certify, rank, or select individuals. In a high-stakes test application, the security issues may be different than the approach taken in e-asTTle; however, the potential exposure of a child’s or school’s name and test scores always represents a threat to the public acceptability of any online system. The lessons learned in asTTle arose in a time that was quite different in terms of information technology features and reliabilities, let alone the much more limited access to laptops and tablets, than can be seen in today’s developed societies. It may be that the lessons learned in asTTle have much greater applicability to developing nations that do not have robust IT infrastructure yet. Not only have technology and infrastructure changed since the advent of asTTle, but also so have teachers’ and students’ familiarity with computers, software, and the internet. That suggests an e-asTTle devised today might look quite different than the current version. Nonetheless, developers and policy makers seeking to transfer from paper-based to online testing would be well advised to consider the issues identified in this paper.

Hence, if governments, societies, and agencies want students to experience online testing as part of their educational experience, it behoves them to ensure that quality and equity are ensured. The e-asTTle experience identifies important challenges with significant cost and infrastructure implications in reaching for this potentially valuable goal.

Notes

Notes on contributor

Prof Gavin T. L. Brown is the Associate Dean Postgraduate (Research) and Director of the Quantitative Data Analysis and Research (Quant-DARE) unit in the Faculty of Education and Social Work at the University of Auckland. He is an Honorary Professor in the Dept of Curriculum & Instruction at the Education University of Hong Kong and an Affiliated Professor in the Dept of Applied Educational Science at Umeå University, Sweden. His research is on educational assessment and its effects on the psychology and behaviour of teachers and students. He is the author of Assessment of Student Achievement (Routledge, 2018).

Disclosure statement

No potential conflict of interest was reported by the author.

ORCID

Gavin T. L. Brown http://orcid.org/0000-0002-8352-2351

References


