From smart testing to smart learning: how testing technology can assist the new generation of education

Susu Zhang*
University of Illinois at Urbana-Champaign, 
603 E Daniel St., Champaign, IL, 61820, USA
Email: szhan105@illinois.edu

Hua-Hua Chang*
University of Illinois at Urbana-Champaign, 
603 E Daniel St., Champaign, IL, 61820, USA
and
East China Normal University, 
3663 North Zhongshan Road, Shanghai, 200062, China
Email: hhchang@illinois.edu
*Corresponding authors

Abstract: This paper provides a brief overview of the theories and research in modern measurement and test theory, also known as smart testing, and further discusses how smart testing relates to the concepts and practices in smart learning. An introduction to smart learning and some of the challenges it is facing will be provided, followed by an introductory survey to a selected few topics in psychometrics, such as item response theory, computerised adaptive testing, large-scale assessments, cognitive diagnosis, and linking, etc. A couple of models regarding the implications of smart testing theories and techniques in smart education will then be proposed, together with the descriptions of some of the ongoing projects combining the two, as well as a discussion on potential future research directions.

Keywords: computerised adaptive testing; CAT; cognitive diagnosis; personalising online learning; mobile learning; smart testing; smart learning; massive online open courses; MOOCs.

Reference to this paper should be made as follows: Zhang, S. and Chang, H-H. (2016) ‘From smart testing to smart learning: how testing technology can assist the new generation of education’, Int. J. Smart Technology and Learning, Vol. 1, No. 1, pp.67–92.

Biographical notes: Susu Zhang is a PhD student in Quantitative Psychology at the University of Illinois at Urbana-Champaign. She received her BA in Psychology from Bryn Mawr College and her BA in Mathematics from Haverford College, Pennsylvania, USA. Her primary research interests are item response theory, computerised adaptive testing and cognitive diagnosis.

Hua-Hua Chang is a Professor of Educational Psychology, Psychology and Education at the University of Illinois at Urbana-Champaign. He is the Editor-in-Chief of Applied Psychological Measurement, former President of the Psychometric Society (2012–2013) and a Fellow of the American
1 Introduction

In the last few decades, a lot of developments were seen in the field of education, aiming to extend learning beyond the traditional classroom experience to further benefit the students. Alongside, the increasing advocacy for individualised, performance-based, globally integrated, and well-rounded learning, theories and technologies for educational assessments, which go far beyond the traditional, basic testing schemes, were developed to provide a solid theoretical framework for educational assessment and to address various needs stemmed from the modern innovations in education. Modern theories in educational assessment are rapidly transforming testing from unaccommodating and stress-creating ranking measures into flexible, reliable, and informative tools that can be used to address the compelling needs of, and hence benefit, various stakeholders in education. This theory-based, trustworthy, flexible, and informative testing is known as smart testing. We discuss, in the present paper, how some of the new developments in testing, such as cognitive diagnostic computerised adaptive testing (CD-CAT), can be used to assist, instead of sabotage, learning. A brief introduction to smart learning, including some of the trending movements, such as individualised learning and massive online open courses (MOOCs), will be provided, followed by an overview of concepts and new developments in educational measurement, and finally, we will discuss how some of the cutting-edge smart testing technologies can facilitate smart learning.

2 Smart learning

2.1 Overview of the concept

The CEO and President of IBM, Sam Palmisano, brought up the initiative of ‘Smarter Planet’ in November 2008 (Palmisano, 2008). It emphasises the potential of cutting-edge technology to achieve global growth and development. Following the introduction of ‘Smarter Planet’, IBM specifically dedicated a report on how smart technology, specifically consumer IT, open source technologies, and cloud computing, can shape the future of education (Rudd et al., 2009). They predicted that how people learn would change significantly in the next decade and listed a few main directions of development, namely technology immersion into students’ learning, flexible and individualised learning paths, knowledge and skills fulfilling the economy’s demands, globally integrated and accessible systems and resources, and leading role of education in the new century.

What IBM has described is sometimes referred to as ‘Smart Education’. A brief definition of it was provided by Zhu and Shen (2013), as creating a smart environment for learning that promotes the development of both the learner’s well-roundedness and specialised competency, which, ultimately, will create value for the entire society.
The key features of smart education are interconnected with one another. To achieve education’s adherence to the economy and its leading role in the society, learners need to develop both the foundational skills required by the service-based economy and the special strengths that are pertinent to the society’s needs. Technology’s integration with education, achieved through any device learning, will improve students’ technological literacy, which is required in today’s economy and will enable the learners to access all kinds of resources, and, together with global integration of systems and resources, it will increase education’s accessibility and availability; the objective is that everyone, no matter if s/he has day jobs, whether s/he is from a privileged or an underprivileged group, and wherever in the world s/he is from, has the chance to receive the same education, and that people can learn in a wider variety of contexts, in terms of time, location, and the media via which they receive their training. Whereas traditional schooling can only be conducted to a limited population in a specific timeframe at a certain location, smart education has the potential to benefit everyone in the society. People who do not have access to traditional education will have the opportunity to receive smart education, and those who already do will receive education designed for a wider audience group and hence increase cultural and societal awareness.

On the other hand, the specialised area for a specific learner should be determined based on a combination of different factors, including the society’s current demand and the learner’s interest and strength. The former requires good understanding of the society and the economy at the time, and the latter relies on individualised educational diagnostics, advisory, and service powered by advanced technology such as learning analytics – the data collection, analysis, summarisation, and presentation centred around learners and their learning environments (Siemens and Baker, 2012; Zhu and Shen, 2013). Traditional classrooms are instructor-centred, i.e., however, large the teacher-student ratio and however similar the students are in terms of performance, a group of students receive the exact same lecture from the instructor. Smart education, however, is student-centred, and every student receives flexible, individualised education based on their acquisition of the material, their strengths, and their passions. Another term for this kind of smart and personalised learning is called adaptive learning. This can hardly be achieved without the immersion of technology in education, and without the basis of globally integrated and open systems and resources, it would be very difficult to implement adaptive learning on a large-scale. Although it still faces a lot of technical limitations, such as limited internet connectivity and speed for mobile devices (Rudd et al., 2009), with the increasing coverage of smart devices and technologies in the world, the advent and soaring popularity of online communities and open digital platforms for sharing information and resources (e.g., Moodle, Wikipedia, and Google Scholar), and the gradually evolving use of digital media in education, adaptive learning is becoming increasingly prevalent, and many pioneers exploring and expanding this possibility emerged and rapidly developed.

2.2 Current developments in smart learning

On June 17, 2013, an article called ‘A is for adaptive’ came out on the Time magazine (Webley, 2013). The article mainly discussed the blooming trend of adaptive, personalised learning in schools. Knewton system, one of the leaders in personalising online learning, enables students to take different learning routes according to different
mastery level of previous knowledge and skills. Two students taking the same course begin with the same materials, but based on their performance, Knewton might direct one of them to a subsequent section that is more difficult, or it might lead one student to a new topic while offering the other more support and supplement for the current section. If we use dots to represent knowledge and skills and connect the dots based on an individual’s learning route, the resulting pictures could be different for two students in the same class. This adaptive nature of instruction allows every student to proceed on his or her own pace, so fast learners do not need to wait for the entire class, and students who have not completely acquired the material can take more time and receive more support.

Khan Academy, a free virtual classroom that offers classes divided into units of ‘microlecture’ videos and interactive tasks, also aims at humanising the classroom by using technology. Teachers may assign lectures to students after school, and students can watch, pause, or replay the lectures as they like. After each lecture, students complete a set of interactive exercise questions to ensure their engagement and to assess their understanding of the knowledge and skills. Correct answer to all the questions demonstrates full acquisition of the materials and serves as the prerequisite for moving on to more complex and advanced materials. Students who have difficulties completing the exercises can re-watch the video lectures, refer to the hints provided for the exercise to find out where specifically they made the errors, or seek support from the online learning forum. Teachers can create their Khan Academy instructors’ accounts as well and receive real-time data on the performance of every student and, instead of giving the same lecture to everyone during class time, offer in-person support to those who get stuck. Also, they can spend class time on more creative and in-depth explorations and discussions (Sinha, 2012).

Online courses by Khan Academy are now widely used in elementary and secondary education, with multiple piloting schools and school districts in the USA and around the world. It is exceptionally valuable in supplementing the instruction of math and science classes. Traditional math and science classes are lecture-based, and what is taught previously usually serve as the foundation of later materials. Some students get stuck on certain knowledge or skills and hence lose track of what is going on later in the class, thus, as the class goes on, we often see students diverging widely in performance. Additionally, spending class time on lecturing limits the amount of time available for discussions and explorations, which makes these engaging in-class activities rarely seen in math and science classrooms, unlike in a lot of humanities and social sciences. The use of technology offers a firm safety net for students when they experience difficulties in understanding some specifics and frees up the classroom, and it was shown to be effective in improving students’ performance, especially in decreasing the proportion of students who score below basic (Sinha, 2011). Up to 2014, Khan Academy attracts more than 10 million unique users every month (Jacobs and Khan, 2014). Similar blended learning and online learning frameworks are extensively deployed in the USA. More than 1 million secondary students take online courses as part of their curriculum each year, and this number is likely to keep going up, as there is the increasing need to provide high quality public education with lower cost (Journell et al., 2014). Depending on when and where the online education is provided and the type of supervision and support received by the students, blended learning model can be divided into different categories, such as the face-to-face driver model, where online modules are occasionally given to students in the classroom to supplement traditional classroom experience; the rotation model, where class-time is alternately spent on instruction and self-paced online learning; and the
From smart testing to smart learning

From smart testing to smart learning

online lab model, where the whole course is given online and teachers and assistants offer individualised support online (Horn and Staker, 2011).

Intelligent tutoring systems (ITS) (VanLehn, 2011), which are computer-based platforms providing interactive exercises to students and step-to-step guides and hints, are also widely adopted to supplement regular education. Based on the ICAP theory (Chi and Wylie, 2014), that learning is most effective if conducted interactively, ITS are created with the intention of facilitating learning through interactive exercises. These tend to be used a lot in Math and Computer Science teaching, because exercises with step-wise, individual need-based guidelines are both feasible to make and effective for these subjects. Some classes on Khan Academy use ITS-based exercises, and many platforms are created and used in education (Koedinger et al., 1997; Schiaffino et al., 2008) and in vocational training (Lajoie and Lesgold, 1989; Eliot and Woolf, 1995).

Another blooming trend in smart education is the prevailing online learning opportunities for higher education. Since Massachusetts Institute of Technology (MIT) first started putting audio lectures and slides on their open course ware (OCW) website in 2001 (Abelson, 2008), numerous top universities around the world, such as Rice University (Connexions Project) and Carnegie Mellon University (Open Learning Initiative), have participated in the open educational resources (OER) movement and made class resources available to the public online (Atkins et al., 2007). In recent years, the idea and practice of OCW and OER, where the focus is mainly instruction and resource provision, evolved into MOOCs, where interactive learning is the focus and is supported by instruction, resource provision, peer evaluation, and certification, etc. (Liyanagunawardena et al., 2013). Exponential growth of customers and available courses has taken place in the MOOC industry. Collaborations have been established between universities who want to make their educational resources public and MOOC organisations who can provide the platforms, such as edX, Udacity, and Coursera, among which, Coursera, the largest MOOC company in the world, has over 10 million unique users and 839 courses provided by 114 institutions available online for free (as of October, 2014) since the debut of the organisation in April 2012 (Larson, 2014). Introduction to artificial intelligence, one of the earliest courses on Coursera, established a paradigm that incorporates instructional videos, automated assessment, peer support, and student-made interactive tools altogether, and this model was adopted by a lot of later courses. Students learn from the lecture videos, practice through quizzes and homework that are either computer graded or peer-graded, seek support from the instructors, teaching assistants, or classmates on the class forum, and, sometimes, generate tools and examples that can facilitate other students’ learning.

With the support of cloud computing, many organisations recently developed learning interfaces for mobile phone, iPad, and other handheld devices. Coursera’s mobile version came out first on iOS in December 2013 and later became available for other operating systems. Students can stream the video lectures, view their course dashboard, and complete the quizzes through their mobile devices using mobile data or wireless network, and they can also download the videos under wireless network for offline viewing later on, without using any mobile data. Similarly, Khan Academy also developed its mobile version, and the most recent update (as of January 20, 2015) made the entire library available on iPad. Hand-written input is supported, and students can access all the contents, including the videos, interactive activities, assessments, and personal profiles. Small chunks of time that are spent on waiting for lunch to be served, on commuting
between work and home, and on trying to fall asleep in bed, can now be utilised to watch microlecture videos. Mobile learning, although still in its beta stage, has further improved the flexibility and accessibility of education.

As mentioned above, online smart learning could not have prospered without the cloud computing technology and globally integrated systems and resources that are distributed to and widely used by the public, such as social networks, forums, communication tools, and digital platforms for information collection (Fini, 2009). At the same time, the large-scale and sustainable development of MOOCs also relies on a nearly self-sufficient operating mechanism.

The professors put their lectures and resources on the MOOC platforms, and in exchange, the MOOC organisations turn their online courses into a licensed product that can be used in the future. And to minimise the cost of assessments and academic advising, automatic grading, peer-assessment, and discussion forums are widely used by major MOOC organisations (Anderson and McGreal, 2012). As part of their requirement for course completion, students usually need to evaluate and provide feedback for several peers’ assignments, with the rubrics and guidelines provided by the MOOC organisations. Assignment scores are calculated based on the peer assessments (Piech et al., 2013). When students have questions, they bring them to the discussion forums. Usually, questions are answered by other students, and the instructors and teaching assistants only need to verify the correctness of the students’ answers and provide responses to the unanswered questions. These approaches keep the cost of MOOC provision at minimum for the organisations. Meanwhile, these organisations and universities also profit from offering certificates upon course completion at a low cost (e.g., Signature Track by Coursera, and Verified Certificate by edX), consented sharing of students’ information with hiring industries and institutions, and, most recently, collaborating with institutions to provide actual credit or degrees, such as the Master’s degree in Computer Science from Georgia Institute of Technology provided by Udacity.

Similar developments of adaptive learning are also seen internationally. A Chinese internet company called NetEase Inc. started its open courseware movement in 2010. Collaborating with many Chinese universities and US-based open courseware organisations, such as Coursera (Penna, 2013) and Khan Academy, it provides free online courses to the Chinese audience. Besides serving as an integrated platform to assess different sources of online courses, NetEase is also adding Chinese subtitles to a large proportion of the course videos. Users can also install their mobile app, downstream the lecture videos, and watch them offline.

Additionally, like in North America, adaptive learning is also starting to integrate with primary and secondary education abroad. One example would be the e-textbook and e-schoolbag project in Minhang District, one of the major districts in Shanghai, China. This project provides a smart learning platform to more than 8,000 students from 160 classes at 80 schools as of 2013 (Zhu, 2014). Students can access their electronic textbook and profile from terminal devices, such as tablets provided by the school or computers, both in the classroom and outside of school. These e-textbooks are designed by e-publishers and content experts and contain interactive, multimedia modules. During class time, teachers can manage what is displayed on students’ devices during classroom and get timely feedback on how well students are doing on in-class exercises and assessments. Outside of the classroom, through broadband or 3G network, students can log on to their e-Schoolbag account to explore the materials more extensively and complete their assignments. The platform supports automatic grading for fixed response
problems, handwritten input, and manual grading, and it also provides summaries of students’ progress that is viewable to the students, instructors, and teachers, with information such as the questions they answered correctly and incorrectly on the assignments. The e-schoolbag supplements learning by providing each stakeholder in education, including the teachers, the parents, and the students themselves, with information on the students’ learning progress, which enables individualised learning paths that are adaptive to the students’ needs. Schools in Minhang district have been experimenting on various adaptive learning models, where the digital modules can be used for entire lectures, for occasional in-class interactions and assessments, for watching recorded lectures outside of class, or for preview, review, and homework.

2.3 Limitations of current practices of smart learning

This exponential growth in the development and usage of smart learning, however, does not prevent it from experiencing many challenges. Although one of the most straightforward significance of smart education is that it lowers the entrance threshold and provides everyone with the access to the resources, to some extent, the audience who can benefit from online learning opportunities might be even narrower. Retention rate of online courses is extremely high, with less than 13% of course completion rate for most MOOCs (Haber, 2013). One of the main reasons is that online open courses have very high requirements on the students. In order to complete a course and actually benefit from it, the learner needs to exhibit high self-regulation, show sufficient motivation, and possess adequate time and mental energy. Most MOOC classes, unlike Khan Academy, try to simulate the traditional classroom experience in higher institutions by setting a course period, posting one-sided lecture videos, and offering old-fashioned homework and exams graded by peers and TAs. This posts a big barrier for the adult learners with additional coursework or career, due to the scarcity of time and mental energy. Additionally, MOOCs face an even wider range of audiences than actual university courses and K-12 online courses. People with different entrance level start with the same video lectures and may lose motivation quickly if their entrance level is either higher or lower than required. Given that a lot of MOOCs are free or at very low cost, the learners’ investment in the courses is low, which is likely to lead to low commitment (Rusbult, 1983). And unlike school-aged children who can be engaged using game-like activities and badges designed by the online learning programs, adult learners tend to be less attracted by these tricks. Therefore, MOOCs usually benefit only the people who possess the time, self-control, high intrinsic interest and motivations in the topic, and the appropriate level of academic background.

The challenge described above is rather specific to MOOCs, but some limitations of smart education apply to most smart e-learning programs. One of those is the ‘digital divide’ described in an article on the New York Times (Ritchel, 2012). The increasing use of digital devices in education can presumably assist with filling the socio-economical gaps in education, provided that the digital devices are properly used for educational purposes. It was shown that students with lower socio-economic status (SES) tend to spend a significant amount of time on social networks and entertainment with the devices. Parents of these children usually have lower digital literacy and awareness about proper use of digital devices, and, according to Lareau (2003), they tend to dedicate significantly less amount of time to guidance and education of children compared to the
more affluent families. As a result, a fault line dividing affluent students who use the
devices for education and underprivileged students who use their devices for recreational
activities is created by e-learning, and the gap across different SES groups gets even
larger.

Some practitioners and researchers (Cappelli, 2014) also argue that the most valued
qualities in companies are usually not hard knowledge but intangible skills, such as
interpersonal communication and collaboration, creativeness, and critical thinking. This
concern goes together with what IBM envisioned as the ‘foundational knowledge and
skills required by the economy’ in their smart education initiative. Traditional classrooms
offer the students with the opportunity to interact with and model their peers and
teachers, and it is unclear whether the same can be achieved through virtual classrooms.
Although some pioneers in smart learning like Salman Khan can argue that e-learning
frees up class time for the development of more intangible qualities in students, and
specialised competency in certain areas is highly valued, how the soft skills can be
incorporated into smart e-learning is still an important question.

Another concern in K-12 adaptive learning is how it fits in to societies with
competitive performance-based personnel selection. For learners, standardised test scores
are commonly used for high school and university admissions, and for education
providers, standardised test scores are usually the criterion for their performance
assessment. Smart learning may allow for more flexibility and room for personalised
development, but both the schools and the students are pressured to perform well on
achievement tests. As a result, it is challenging for the schools to make large adjustments
and take full advantage of adaptive learning, and students receive heavier burdens outside
of the classroom, because they need to engage in self-regulated learning and keep up with
what was covered in the traditional curriculum at the same time. Given the scarcity of
time and resources for the students and the schools, it is hence necessary for adaptive
learning to find its niche in the current educational institution and assist learning in the
most efficient way.

In the following sections, we will discuss how some of these limitations of smart
learning can be addressed using smart testing, especially adaptive testing. Adaptive
testing, also known as tailored testing, adapts the difficulty of the questions to the
examinees’ ability level estimated using previous responses. The field of psychometrics
and some of the popular research topics will be briefly introduced, followed by a
discussion on how testing theories and technology can assist learning. We will further
introduce some of the ongoing projects and real-world examples in the USA and
worldwide. Lastly, we will ruminate on and discuss some possible future directions of
research on this topic.

3 Smart testing

The concept and characteristics of smart testing might be better explained if we make
contrasts between smart testing and traditional testing. Here, traditional testing refers to
the linear paper-and-pencil tests administered to small, specific groups, and the total
score is the only index used to summarise an individual’s performance. This type of
testing is backed up by classical test theory (CTT). The most important characteristics of
smart testing are presumably its theoretical soundness, flexibility, trustworthiness, and
the potential to be used for various purposes. More specifically, it should be backed up by
solid theories, it should provide generalisable and reliable measurements for the entire population of interest, the questions given to each subject should be individualised, and the results of the test should be informative for purposes other than ability comparison. Modern test theories, such as item response theory (IRT), provide the theoretical basis for smart testing. To understand the limitations of traditional testing as well as the advantages and challenges in smart testing, we will provide a brief introduction to the Psychometric theories behind them.

### 3.1 Psychometrics and IRT

Psychometrics is the study of methodologies for measurement, including measure design, administration, and analysis, of people’s psychological constructs, such as personalities, attitudes, and abilities (Chang and Ying, 2007). In the context of educational assessment, the underlying trait of interest is usually individual’s abilities, in terms of knowledge and skills, in the designated field. Sub-domains under psychometrics include CTT, IRT, generalisability theory (G theory), etc.

#### 3.2 Classical test theory

CTT is presumably one of the simplest theories in psychometrics. The foundational model for CTT states that the random variable for the observed score of an examinee on a test, $X$, is the sum of the examinee’s unobserved true score, $T$, and error, $E$. In other words, $X = T + E$. The expectation of error, $E(E)$, is 0, and the expectation of observed score, $E(X)$, is the individual’s true score $T$. Therefore, observed score, computed as the sum of item scores, is an unbiased estimate of the individuals’ true score. The variance of error, $o^2$, is assumed to be the same across participants, meaning error is assumed to be unrelated to true score (Allen and Yen, 2001). With a few other assumptions, one can investigate many properties of a specific test and the items involved, such as item difficulty, item discrimination, test reliability, and test validity. One of the limitations of CTT, though, is that the results are not generalisable beyond a specific test, that is, the true score, $T$, in the model only captures the true score on a test. The properties of the test, hence, are population- and context-specific. For example, item difficulty is captured by the proportion of examinees who answer an item correctly. If we give the same item to another set of test-takers, the difficulty level for the same item will vary. For this reason, one cannot make predictions on the performance of individuals on an item unless the item was previous administered to a similar population (Lord, 1980; Hambleton, 1991).

#### 3.3 Item response theory

IRT, on the other hand, defines each item by a set of item parameters and each examinee by examinee parameters, and it uses an item response function (IRF) to describe the relationship between participants’ responses and the item and examinee parameters. The most commonly seen examinee parameter would be $\theta$, the latent trait being measured, such as personality, ability, or other psychological constructs. For a dichotomous item, where the outcome is either correct (1) or incorrect (0), an IRF establishes the regression of probability of correct answer on $\theta$ and the item parameters. One of the commonly used
IRF’s is the three parameter logistic (3PL) model. For a given ability level \( \theta \), the probability of correctly answering item \( j \) is

\[
P(X_j = 1 | \theta) = P_X(\theta) = c_j + (1 + c_j) \left[ 1 + \exp \left( -1.7 a_j (\theta - b_j) \right) \right]^{-1},
\]

where \( b_j \) is the item difficulty, \( a_j \) is the item discrimination, and \( c_j \) is the pseudo-random, or guessing parameter. In the special cases where all items have the probability of guessing correctly \( c_j = 0 \) and where all items have item discrimination of \( a_j = 1 \) and guessing parameter \( c_j = 0 \), the 3PL model is reduced to the 2PL and the 1PL models, respectively. Other promising models that work well in different contexts also exist, such as Masters’ (1982) partial credit model for polytomous responses, Bock’s (1972) nominal response model that treats different incorrect responses differently, etc. Conventional IRT models for ability testing usually have two underlying assumptions (Hambleton, 1991). The first is monotonicity of the IRF. In other words, given any item \( j \) and any two ability levels \( \theta_1 > \theta_2 \), probability of answering the item correctly should be higher for the examinee with higher ability, i.e., \( P(X_j = 1 | \theta_1) > P(X_j = 1 | \theta_2) \). The second assumption requires local independence of item responses for a given ability level, meaning that for a given \( \theta \) level, probability of answering an item correctly does not depend on whether the previous item(s) were answered correctly. Let \( X_1, \ldots, X_n \) be a series of items administered to an examinee with ability \( \theta \), the local independence assumption requires that \( P(X_1 = x_1, \ldots, X_n = x_n | \theta) = P(X_1 = x_1 | \theta) \cdots P(X_n = x_n | \theta) \). These two assumptions are substantial for estimation and analysis of the data under the IRT framework. Therefore, before conducting any subsequent data analysis, we should always check the monotonicity of empirical item characteristic curves for each item (plot of proportion of people who answered the item correctly against total score on the test), and the dimensionality of \( \theta \) should be determined for local independence to hold. The major advantage of IRT over CTT is that by partitioning the item effect and participant effect, under a model that fits the data appropriately, the item parameters will remain invariant across populations and the ability parameters remain invariant across tests. This enables us to put the examinees’ ability level on the same scale, even if the sets of questions they answered differ. Later, we will also see that IRT serves as an important basis for computerised adaptive testing (CAT).

### 3.4 Adaptive testing

The practice of adaptive testing first emerged in the context of sequential test for dose-response relationships, where various levels of stimulation would produce a dichotomous outcome. An example would be finding the lethal dosage (LD50) of a drug through animal studies. Researchers are interested in finding the dosage of a drug that has 50% chance of killing the animal. Robbins and Monro (1951) proposed stochastic approximation as a method to find this critical dosage by adjusting the amount administered in each trial. The problem is reformulated as a root finding problem: For constant \( \alpha \) and a monotonically increasing unobservable function \( M(x) \), the equation \( M(x) = \alpha \) has solution at \( x = \theta \). Since no analytic form is available for \( M(x) \), an unbiased estimator for all levels of \( x \), \( N(x) \), is used for obtaining measurements on \( M(x) \). In the case of finding the median lethal dose (Cochran and Davis, 1965), \( M \) is the underlying function mapping a dose level to the probability of death, and we try to find a level of
dose, \( x_0 \), such that \( M(x_0) = 1/2 \). We estimate \( M(x_k) \) at dosage level \( x_k \) with the proportion of animals administered with \( x_k \) amount of drug who died, denoted \( p_k \). Then, given the dosage in the \( n \)th trial, \( x_n \), the dosage to be administered in the \( (n + 1) \)th trial can be chosen such that:

\[
x_{n+1} = x_n - a_n \left( p_n - \frac{1}{2} \right),
\]

where \( a_1, \ldots, a_n \) are predetermined decreasing positive step sizes. Robbins and Monro (1951) showed the asymptotic convergence of \( x_n \) to the critical dose \( x_0 \) as \( n \to \infty \). Although eventual convergence is guaranteed under certain reasonable assumptions, fast convergence to the critical value is desired, and this can be achieved by effective choices of step sizes.

Lord (1971) extended the idea to the field of measurement to develop tailored testing, which is what we later would call adaptive testing. The overarching goal, instead of administering a uniform set of test items to each examinee, is to choose items for a given examinee based on his/her performance on the previous questions, so that accurate measurement of the individual’s ability can be obtained with a small amount of items. To do this we need to have a large, extensive item bank with items at different difficulty, discrimination, and guessing levels. For this reason, IRT serves as an important theoretical basis for adaptive testing. Under the 3PL model, when an item with difficulty of \( b_0 \) is administered to an examinee with ability \( \theta \), we can plug it into the IRF and see that the probability of answering the item correctly will be \( 1/2 \). Thus, the problem of finding the examinee’s ability \( \theta \) can be reframed into a sequential test problem of finding the item with difficulty \( b_0 \) such that the conditional probability of answering the item correctly, \( P(X | \theta, b_0) \), is equal to \( 1/2 \). To maximise speed of convergence, Lord (1980) proposed the maximum information approach for item selection, where the next item \( j \) is chosen so that, given the current ability estimate \( \hat{\theta} \), the Fisher information,

\[
\frac{\log f(b_j)}{\log f(\hat{\theta}) - \log f(\theta)}
\]

(Birnbaum, 1968; Lord, 1980), is maximised. The inverse of the Fisher information is the lower bound of variance of any unbiased estimator for \( \theta \). Therefore, by maximising FI, we minimise the lower bound of \( \text{Var}(\hat{\theta}) \), which hence maximises efficiency. The information function of an item with respect to \( \theta \) usually takes a bell shape, with the centre at its difficulty level and height positively related to its discrimination parameter. Given two items with the same difficulty \( b_i = \hat{\theta}_{n-1} \), where \( \hat{\theta}_{n-1} \) is the current ability estimate for the examinee, the item with higher \( a \) will have higher Fisher Information. On the other hand, given two item with the same discrimination, the item whose difficulty is closer to the current ability estimate \( \hat{\theta}_{n-1} \) will have higher FI.

Adaptive testing has prospered with the increasing prevalence of computers, and compared to traditional linear tests, by selecting items that pinpoint the ability of the test-taker, CAT provides more accurate ability estimates with much smaller amount of items (Weiss, 1982), which reduces examinee’s burden and reduces the risk of cheating (Hambleton, 1991). For this reason, CAT has been adapted in many high-stake exams, including the graduate management admission test (Rudner, 2010), the armed services vocational aptitude battery (Segall et al., 1997), and the nurse licensure exam (NCLEX) by the National Council of State Boards of Nursing (Beeman and Waterhouse, 2001). The maximum Fisher information item selection method proposed by Lord (1971) is presumably the most commonly used item selection method for aptitude testing, but it has
some non-negligible setbacks. For example, at the early stage of the test when the ability estimate \( \hat{\theta} \) is unlikely to be accurate, the maximum Fisher information approach using current ability estimate \( \hat{\theta} \) as an estimate for \( \theta \) does not perform optimally and may lead to divergence. To address this issue, Chang and Ying (1996) proposed selecting the next item \( j \) that maximises the Kullback-Leibler index (KI), defined as follows:

\[
KL_J(\hat{\theta}) = \int_{\hat{\theta} - \delta_n}^{\hat{\theta} + \delta_n} KL_J(\hat{\theta} \parallel \theta) d\theta,
\]

where \( \delta_n \) specifies the interval of integration, and \( KL_J(\theta \parallel \theta) \) is the Kullback-Leibler information (Kullback and Leibler, 1951), given by

\[
KL_J(\hat{\theta} \parallel \theta) = E_{\hat{\theta}} \left[ \log \frac{P(x_1, \ldots, x_n \mid \hat{\theta})}{P(x_1, \ldots, x_n \mid \theta)} \right].
\]

KL information is the expectation of the log-likelihood ratio between the two distributions, so by integrating it over a pre-specified interval around \( \hat{\theta} \), the KL index measures the discrepancy between the distribution of the current ability estimate \( \hat{\theta} \) and the true \( \theta \) within the interval. Unlike FI which only uses information at the point \( \hat{\theta}_n \), when \( \delta \) is large, the KL index uses information within a large neighbourhood of \( \hat{\theta}_n \), so it is a global information measure and performs better than the FI local information at the beginning of the test, when \( \hat{\theta} \) cannot be accurately estimated. It should be noted that the FI is proportional to the second derivative of KL information. Since the FI is more efficient when \( \hat{\theta}_n \) is relatively close to \( \theta \), at later stage of test, FI should be used, and this is equivalent as using KL-index by taking small \( \delta \).

Another limitation of the maximum Fisher information approach is that it often tends to select items with high discrimination, and as a result, a particular set of items are overexposed while others might never be administered. This is problematic for two major reasons. First, although controlling the difficulty level of items during item writing is not too difficult, it is practically unfeasible to purposefully write high discrimination items. The common practice of test developers is to create a large set of items, obtain item parameters based on pilot tests, and only keep the items with relatively high discrimination. Developing an item bank with high discrimination is hence very difficult. Second, excess exposure of a small set of items menaces test security, as the risk of someone memorising and exposing the item to others is larger. The a-stratified method (Chang and Ying, 1999) provides a solution to this by administering low discrimination items at the beginning of the test, when the ability estimate might not be close to true ability, and later on in the test, items with high as are selected to achieve high precision.

Stemmed from the blooming of IRT and CAT, a lot of questions related to making testing more efficient, accurate, practically feasible, and flexible, have emerged. These questions have been extensively researched on, and a selected few are described below. Smart testing has also been used in many other contexts, such as personality measurement (Stark et al., 2012) and diagnosis of psychological disorders (Gibbons and Hedeker, 1992). Here, we focus on the context of ability measurement.
3.4.1 Testlets
The local independence assumption in IRT can be violated if certain items in the test are testing on the same issues. Testlets are commonly seen in language tests, where a series of questions may follow from a single reading comprehension paragraph or listening exercise. Using traditional IRT models assuming local independence may lead to biased estimation. Early researchers proposed treating a testlet as a single polytomous item and analyse it using graded response models (Thissen et al., 1989). Later, based on the 2PL and 3PL, testlet response models (TRM) accounting for the dependency structures between items were developed (Wainer et al., 2000) and further increased measurement accuracy.

3.4.2 Differential item functioning
One of the key underlying assumptions of IRT is the invariance of item parameters across populations. This assumption is not warranted and consistently needs to be checked upon. Differential item functioning (DIF), also known as item bias, refers to the case where the same item has different parameters across populations. An example would be an item related to probability theory in the context of a baseball game. This item would be more difficult for individuals who grew up in another culture where baseball is not a prevalent sport. There might also be DIF across two groups who took the test at different time points. For example, an item may be a lot easier for a group who took the test later than those who did earlier, which suggests that the item might have been leaked. These items would threaten the fairness of a test and need to be deleted. A comprehensive collection of literatures on DIF detection can be found in Holland and Wainer (2012). Besides the implications in assuring test fairness, detecting DIF and ensuring measurement invariance are highly related to smart learning. As smart learning advocates for globalised platforms for education and resource sharing, globalised educational assessments should also be developed to measure the learning progress for students worldwide. Assessments along this line, such as the Program for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS), have already been developed and used in many cultures. To accurately assess the progress of students in different cultures, test items that are biased against or in favour of specific cultures need to be detected and revised.

3.4.3 Linking and equating
Although the purpose of a test is to measure the ability of the test takers, the item parameters also need to be estimated. Thus, we usually need to simultaneously estimate the item parameters and the examinees’ ability level at the same time based on the responses. A problem with such joint estimation is the scale indeterminacy, meaning that the resulting estimates are unique only if the scale is fixed (Hambleton, 1991). To be able to proceed with the estimation, we must fix the scale of either the ability estimate or the item parameter estimate, usually by standardising the estimates on one of them. This scaling needs to be taken into account when we compare the item and ability parameter estimates across multiple groups. Therefore, to put the item (or ability) parameter estimates across groups on the same scale, linking and equating need to be done. One practice is to include a set of common items in the two (or more) tests being equated.
S. Zhang and H-H. Chang

(Klein and Jarjoura, 1985; Vale, 1986), so that these items can serve as the anchors (or reference points) when the parameters estimated from different tests are compared. For global or nation-wide large-scale assessments, in order to ensure score comparability across different administrations of the test, linking and equating can be crucial. Only then can we compare the scores of a student in Asia to another student in North America who took the test on a different day with a different set of test questions.

3.4.4 Content constraints

In CTT, the content validity refers to the extent to which a test addresses all facets of a given construct (Allen and Yen, 2001). A high school math exam should cover all the major content areas, such as trigonometry, statistics, and functions. The maximum information approach in CAT does not place any content constraints on the items selected, so content balancing methods are needed to meet the practical needs of a test, such as having a specific number of items in each content area. Cheng and Chang (2009) developed a maximum priority index (MPI) item selection approach for multiple content constraints.

3.4.5 Multistage testing

Multistage testing (MST) is a special case for CAT where the questions that are adaptively administered to the examinee come in sets, instead of single items. Fully sequential CAT, where each item is adaptively selected, uses fewer items and produces accurate measurements, but it is computationally heavy and impractical in certain circumstances with limited resources, and at the early stage of a test, the ability estimate using a set of items tends to be more accurate than that provided by a single item. For these reasons and with the automated test assembly technology that assembles multiple parallel test forms with specified psychometric properties (Luecht, 1998), MST is now commonly used by practitioners and large testing companies, such as educational testing service. Hybrid designs that combine CAT and MST were also studied to incorporate the strengths of both. Wang et al. (2014) proposed a variable-length MST method called on-the-fly MST that administers larger sets of items to the examinee at the beginning of the test, decreases the number of items in a stage as the test proceeds, and switches to single item adaptive design in the later part of the test for the most precise measurement. This design performed more accurately and efficiently compared to MST and fully sequential CAT.

3.5 Using CAT for learning

Whereas usually tests are conducted to assess examinees’ ability level, there are situations in which we are interested in how well the individuals have acquired certain knowledge and skills. In the context of education, the latter is known as criterion-referenced assessment (Cronbach, 1949). The purpose of criterion-referenced assessment is not to rank the examinees and see their relative standing in the population, but to provide feedback either to the test-takers or the educators on the individual’s progress. This information can assist the individuals in future learning by informing them of their strengths and weaknesses, and it can also be used by educators to adjust teaching strategy to pinpoint individuals’ needs.
3.5.1 Multidimensional IRT and CAT

Multidimensional adaptive testing (MAT) can be used to address both purposes. In the original IRT model, the ability parameter, $\theta$, was designed to be a unidimensional parameter, and a multidimensional generalisation of the 3PL model extended it to situations where multiple skills are assessed in a single test (Reckase, 1985). The ability vector for a given individual $i$ is $\theta_i = (\theta_{i1}, \cdots, \theta_{ip})^T$, where $p$ is the dimension of $\theta$, or the number of sub-areas, and $\theta_{ik}$ is the individual’s ability in the $k$th area. The areas can be dependent on one another and a covariance structure can be specified to account for the relationships between different areas. For the item parameters, as an extension of the 3PL model, the guessing ($c$) and difficulty ($b$) parameters remain unidimensional, whereas the discrimination ($a$), a measure of how much the probability of correct response will change as a result of unit increase in ability, is now a discrimination vector with $p$ entries.

For the $i$th item and the $i$th examinee, the IRF is given by

$$P_j(\theta_i) = c_j + (1-c_j)\left[1+\exp\left(a_j^T \theta_i - b_j\right)\right]^{-1}.$$ 

Note that a low value on one of the dimensions can be compensated by a high one in another dimension. This is known as a compensatory model (Rupp et al., 2010). The relationship among the dimensions can also be non-compensatory, where a low value on one variable cannot be compensated by another, the deterministic input, noisy ‘or’ gate (DINA) (Junker and Sijstma, 2001) model, which we will talk about later, is an example of which. Right now, we focus on the multidimensional 3PL compensatory model.

Segall (1996) proposed maximising the determinant of the Fisher information matrix as a method for item selection in MAT, known as the D-optimal approach. Like in the univariate case where we want to minimise the variance of $\hat{\theta}$ and make the confidence interval of $\theta$ as small as possible, in the multidimensional case, we want to minimise the area of the confidence ellipsoid for $\theta$, which can be done by maximising the determinant of the FI matrix. Through MAT, a regular test can be combined with a diagnostic one, so that not only does the test-taker receive an ability estimate that can be used for personnel selection or other purposes, s/he can also obtain information on the performance in different areas, which can be used for seeking remedial interventions. A KL item selection method in MAT was proposed by Mulder and van der Linden (2010), and the index involves multiple integration within the $p$-dimensional neighbourhood of $\hat{\theta}$, which can be computationally heavy. Wang and Chang (2011) showed that this KL-index is proportional to volume of the sphere spanned by the discrimination, $a$. More specifically, the KL-index

$$KL \approx \frac{2^{p-1}}{3} \left(\frac{r}{\sqrt{n}}\right)^{p+2} \left[1-\frac{P(\theta)\left[P(\theta)-c\right]}{P(\theta)(1-c)}\right]^2 \sum_{k=1}^p a_k^2.$$ 

And using this index for item selection is equivalent to using the multidimensional KL information, making the computation a lot easier.

3.5.2 Cognitive diagnosis

Cognitive diagnosis (CD) models, or diagnostic cognitive models (Rupp et al., 2010), are restricted latent class models describing the relationship between underlying cognitive
profiles, items, and responses. The underlying cognitive profile of an individual is, like in MIRT, multidimensional, but there are a few differences between the latent cognitive profile in CD and the latent ability $\theta$ in MIRT. First, whereas in MIRT typical dimensions are subareas, in CD, different dimensions of the cognitive profile represent different unitary attributes used to solve questions. And second, whereas $\theta$ in IRT is continuous and numerical, the latent cognitive profile in CD is categorical (either dichotomous or polytomous). Examples of attributes in the context of an elementary school math assessment can be addition, subtraction, multiplication, and order of operations. The number of attributes being assessed in a given test and the relationship between the items and the attributes are determined a priori. This information can be summarised by a Q-matrix with rows as items, columns as attributes, and entries specifying whether the item measures the attribute or the degree to which the item measures the corresponding attribute. Before the test, the Q-matrix is usually formed by a group of educational experts based on their judgements, although recent researchers (De La Torre, 2008) also came up with empirical methods for estimating the Q-matrix. On the other hand, a test-taker can have a certain degree of mastery for each of the attributes, forming his/her cognitive profile or mastery pattern, and the purpose of the cognitive diagnostic assessment is to estimate the mastery pattern of each test-taker, based on the Q-matrix, the subjects’ responses to the items, and the model. Individuals’ mastery patterns can be summarised in an A-matrix, with rows as test-takers, columns as attributes, and entries describing whether the individual has mastered the attribute (1) or not (0) if the attributes are dichotomous, or the individual’s degree of mastery for the attribute if the attributes are polytomous.

An extensive amount of CD models have been developed, such as the knowledge-space model (KSM) (Doignon and Falmagne, 1999), the rule-space model (RSM) (Tatsuoka, 1983), the deterministic input, noisy ‘and’ gate (DINA) model (Junker and Sijtsma, 2001), and Bayesian inference networks (BIN) (Levy and Mislevy, 2004). The DINA model is commonly used in educational assessments. It is a non-compensatory model with dichotomous responses and dichotomous latent attributes. Under the DINA model, the probability of answering the $j^{th}$ item correctly by the $i^{th}$ examinee with attribute patterns $[\alpha_{i1}, \ldots, \alpha_{iK}]$ is

$$P(X = 1 | \eta_j) = (1 - s_j)^{\eta_j} g_j^{1 - \eta_j},$$

where $\eta_j = \sum_{k=1}^{K} q_k^{a_k}$, $q_k$ is 1 if the item measures the $k^{th}$ attribute and 0 if it does not, $s_j$ is the probability of making an error on the $j^{th}$ item even when the attributes requires for the item are all mastered, called the slipping parameter, and $g_j$ is the probability of answering the item correctly even when not all required attributes are mastered, called the guessing parameter. From the response data and based on the model and the pre-specified Q-matrix, the item parameters (i.e., the $s$ and $g$ for each item) and the examinee parameters (i.e., the examinee’s latent class/cognitive profile) can be measured. Research has shown that, with a correctly specified Q-matrix, the DINA model can provide accurate estimation of the item and examinee parameters (De La Torre, 2008; Cheng, 2009). While multidimensional IRT can indicate an examinee’s performance in different areas, CD measures the test-taker’s acquisition of more fine grained skills and can help pinpoint the individual’s specific cruxes preventing them from fully understanding the materials.
CAT for CD was later developed, with the aim of accurately and efficiently identifying the test-taker’s cognitive profile. Numerous item-selection approaches have been proposed. Tatsuoka (2002) came up with the method of minimising Shannon entropy, an index measuring the uncertainty with a distribution. Xu et al. (2003) introduced the KL-index in CD context, where the sum of the KL information between $\alpha$ and $\hat{\alpha}$ across all possible patterns is taken in place of the integral over a neighbourhood around $\theta$ in IRT. Cheng (2009) then proposed the posterior-weighted KL (PWKL) approach, which extended the KL-index to a weighted sum of the KL information over all patterns, where the weights are determined by the posterior probability of the patterns, increasing the measurement precision and efficiency. Dual-objective CD-CAT that simultaneously measures ability $\theta$ and attribute profile $\alpha$ were also developed. McGlohen and Chang (2008) suggested choosing items based on both $\hat{\theta}$ and $\hat{\alpha}$ using a shadow test method. The basic idea is to use $\hat{\theta}$ to construct a shadow test (van der Linden, 2000), a test form that contains the $n-1$ previous items and the optimal potential subsequent items satisfying the content constraints, and then use $\hat{\alpha}$ to select the $n^{th}$ item with the KL or the SE indices. This enables incorporation of content balancing requirements into CD-CAT through formation of the shadow tests. Wang et al. (2014) further proposed selection of items based on both $\hat{\theta}$ and $\hat{\alpha}$ through weighted sum of the transformed KL-indices of both $\hat{\theta}$ and $\hat{\alpha}$, enabling accurate estimation of both the ability and the underlying cognitive profile.

4 Combining smart testing and smart learning

In the previous section, we briefly mentioned how modern testing theories and methods can be used to improve the fairness and test- and score-generalisability of large-scale assessments, which is pertinent to the smart learning ideals of developing global platforms for educational resource sharing and promoting equity. In this section, in a different angle, we will look at the potential contributions of smart testing to individualised, tailored learning.

4.1 Motivation

Since the enactment of the No Child Left Behind Act (NCLB) in 2001 in the USA, almost all states joined the Common Core State Standards Initiative (2012) and are mandated to track students’ progress based on annual educational assessments, and the progress of each school on annual assessments is used to determine the school’s funding for the subsequent year. Two large consortiums, namely PARCC and smarter balanced (Herman et al., 2013), were formed to conduct annual educational assessments on how well every student and school meet the common core standards. Many computed-based assessments also stemmed from this, and up to 2009, more than half of the states in the USA have conducted pilot projects on online educational assessments (Quellmalz and Pellegrino, 2009). In 2009, the US federal government initiated the ‘race to the top’ (RTTT) fund, aiming at encouraging and rewarding States making substantial progress in K-12 student achievements, such as increasing overall achievements, closing the gaps between stronger and weaker groups, and promoting students’ preparedness for future
education and occupations. To close the gaps between lower and higher performing students, different interventions should be given to different groups of students based on their individual needs, so higher performing students can explore more, and lower performing students can catch up timely (Chang, 2012).

With the ongoing smart learning movement, many schools have adopted the adaptive learning technology, such as lecture recordings, interactive tutoring systems, and microlectures, and made them available for the students after class. However, as we have discussed earlier, having a large amount of choices and resources can be overwhelming, leading students to receive heavier burden or again abandon these resources due to difficulty in choosing. On the other hand, CAT and CD-CAT can identify individual’s mastery patterns and narrow down the resources to ones that are specifically tailored for the student, so that adaptive learning resources can be used in the most efficient manner.

Some schools and states in the USA are also interested in conducting formative assessments to understand students’ mastery of the materials, which can be used for future curriculum planning. CD-CAT has the advantage of providing rigorous but short assessments of the students’ learning progresses, so it can be administered frequently to the students, so as to offer diagnostic information on each individual student’s knowledge acquisition, which can be used to fine-tune teaching and mentoring.

4.2 Establishing a model combining adaptive testing and adaptive learning

Here, we propose two possible models for using adaptive testing to facilitate adaptive learning. The models are not unique and can be flexibly adjusted based on the teaching modes and the types of contributions testing purports to make.

Figure 1 illustrates how adaptive testing, specifically diagnostic assessments, can be incorporated to a teacher-based classroom. Students periodically receive adaptive diagnostic assessments, and with very little time and few items, a good estimate of each student’s mastery level of the materials can be obtained. Diagnostic report for each student can then be generated and given to the students and the teachers. Students can better recognise the skills they have mastered and ones they need to work on, and teachers can see each individual’s progresses and challenges, as well as the class’ overall progresses. Based on these results, teachers can adjust their lecturing focuses to target the class’ needs and provide personal support to individual students based on their specific challenges and potentials. Previous diagnostic results can also be used to adaptively select items in subsequent tests, so that the number of items assessing skills a student has already mastered can be minimised, and items on previous weaknesses of the student can be given to see the student’s progress.

Figure 2 demonstrates how adaptive testing can be used to select the materials given to the students when instructions are given mainly through an online platform, such as MOOC, e-textbook and e-schoolbag systems, etc. Similar to teacher-based instructions, students are given diagnostic tests from time to time, and diagnostic classifications are generated afterwards. These data, as well as the student’s historical data, can be inputted into an adaptive engine to selectively construct the student’s individual learning platform, with personalised learning routes selected from the whole pool of lecture segments, interactive examples or supplemental materials on selected topics. And like for teacher-based classrooms, performance on previous assessments can be used for designing future diagnostic assessments. Students can also view their performance on the
assessments on the online platform, through which they can learn about the skills they have learned or the ones they have not, as well as their progress over time.

**Figure 1** Adaptive testing and adaptive learning model, where teachers are the main instructors

![Diagram of Adaptive Testing and Adaptive Learning Model](image1)

**Figure 2** Adaptive testing and adaptive learning model, where instructions are mainly given through online platforms

![Diagram of Adaptive Testing and Adaptive Learning Model](image2)

### 4.3 Pioneering Pilot Studies

With the current browser/server architecture, it is very feasible to implement CAT and CDCAT on a large scale. A group of students in Dalian, China took an English proficiency CDCAT in 2011. They participated in the assessment over a three-day period, with more than 2,000 completing the test simultaneously. Liu et al. (2013) also recently carried out a large-scale achievement test in China using CD-CAT version of the Level 2...
English Achievement Exam for 6th grade students. Based on the test blueprint, 400 items and a Q-matrix with eight attributes were created by 11 content experts. To verify and revise the Q-matrix, a group of 6th grade students were asked to complete the items and indicate what skills they used for each item, and six English teachers provided their thoughts on potentially missing attributes. To obtain the item parameters, parallel test-forms were constructed and were administered to a total of 38,600 students, and the 3PL item parameters (a, b, c) as well as the slipping and guessing parameters under that DINA model were estimated. Problematic items were also deleted. The CD-CAT was then constructed, with Shannon entropy as the item selection method, fixed test-length of 36 items, and the final estimate of ability and cognitive profile using Bayesian maximum a posteriori (MAP). After checking the proper functioning of the test through a simulation study, it was then delivered through the school computers connected to the test server and completed by 584 5th and 6th grade students in Beijing. 90 of which were selected for a validation study examining the consistency of performance on the CD-CAT and that on an earlier achievement test. Results indicated that individuals with more estimated mastered attributes also tended to show better performance on an earlier test. Three teachers of these students also evaluated the cognitive profile estimated by the CD-CAT and indicated high agreement with the CD-CAT diagnosis for around 75% of the students.

One thing to note about CD-CAT is that it can be used not only for assessment purposes, the questions themselves can also be good exercise for the students’ reinforcement of the knowledge and skills. By giving different students different sets of items appropriate to their mastery of the knowledge, CD-CAT can also be used as an after-class learning tool that offers students individualised treatments. CD-CAT projects were carried out in Zhengzhou, where students use e-learning systems to complete their individualised assignments at home. Not only did more than 90% of the students indicate that CD-CAT has greatly assisted their learning, teachers also reported that the CD-CAT encouraged students to think independently and offered helpful diagnostic information on the student’s strengths and weaknesses.

The use of CD-CAT should not be limited by the lack of computers. In Dalian, China, students were given booklets of assignments or assessments (Chang, 2015; Wang, 2013). And after completion, their answers were read by a Scantron and transmitted to a computer, where each individual student’s responses can then be analysed and their cognitive diagnostic information can be obtained. Based on the students’ performance, the computer would assign different booklets of questions to different students as subsequent assignment or assessment, and each student received a set of tailored questions. The idea is similar to MST, where questions are adaptively administered in sets. Their use of Scantron and paper-and-pencil adaptive test suggests the possibility of smart testing in the relatively traditional teacher-based classrooms and in schools without the resources to get computers for running the computer-based test.

CD-CAT can be used not only for primary and secondary education but also for higher education. At this present moment, an NSF project is being conducted at University of Illinois at Urbana-Champaign on the use of CD-CAT to assist students’ learning in Science, Technology, Engineering, and Mathematics (STEM) fields and to improve STEM retention. In recent years, a lot of drop-outs from the STEM fields have occurred, and part of it was likely due to the students’ lack of ability to pinpoint their weaknesses. The failure to seek timely remedy leads to unfilled gaps in learning, and subsequent materials built on the previous contents will be harder for the students to
understand, resulting in a positive feedback loop that makes it increasingly hard to catch up. Therefore, CD-CAT might be helpful to help the students quickly identify their issues, which they can use to seek individualised help. About half of the students in a 200-level physics class who are showing some difficulties (with previous exam score of less than 70%) periodically take CD-CATs, and they receive diagnostics reports that can subsequently be used to study by themselves or receive video lecture instructions or interactive exercises on their weaknesses. Preliminary results on the effectiveness of CD-CAT and different individualised treatments are very optimistic, and more data will be collected, examined, and reported in the near future.

4.4 Future research directions

This use of CD-CAT in higher education leads us to think about the possibility of using adaptive testing in online courses, such as MOOCs. These courses tend to attract adults who are already full-time students or people in the work-force. They have drastically different entrance levels, and time is a very scarce resource for these audiences. Ideally, with the help of CD-CAT, individuals’ level of acquisition of the materials can be determined, and individuals with different cognitive profiles can take different learning routes, so that only a specific set of videos lectures are administered to each individual. Current CD-CAT technology, however, has some limitations and cannot fully and flexibly support these purposes. One limitation is that CD-CAT is still relatively limited to questions with fixed responses. Although it can also be used for free-response questions, extensive amount of human labour will be required to conduct the grading. Given that most MOOCs are offered to the audience at no charge or very low charge, a lot of them currently deploy a peer-grading mechanism and use BIN or other graphical models to model the relationship between peer grades, peergrading reliability, class performance, and true score of the individual (Piech et al., 2013). To some extent, this is analogous to the CTT models in Psychometrics. Future research should investigate how CAT and CD-CAT can be more flexibly used for free response questions.

Another challenge is the development of item banks. Item bank development can be very expensive, and given MOOCs self-sufficient nature, spending a tremendous amount of resources on item bank development is rather unrealistic. However, for a lot of the repeatedly offered courses, data on previous audiences’ responses to the questions can be used to obtain the item parameters and the empirical Q-matrix using data-driven approaches, and new items can be incorporated to the item bank using the methods for online calibration in the field of CAT, where new items are blended into existing assessments and are adaptively administered to the appropriate subjects, so that the item parameters can be obtained accurately without large pretest subject pools (Makransky and Glas, 2014).

4.5 Concluding remarks

The advent of smart learning has torn down lot of barriers that were present in traditional education, making education more accessible and flexible. Alongside, the bloom of smart learning was the rapid development of smart testing theories and technologies. Testing and measurement research will see a lot of new potential research directions by
incorporating the needs in smart learning, and application of smart testing will be able to make smart learning more efficient, reliable, and mature.

References
From smart testing to smart learning


