Assessment, Psychometrics, and Technology: State of the Art

White Paper: June 2017

Nathan Thompson, PhD
Assessment Systems Corporation

This paper was published as part of conference proceedings by Kuban State University, Russia. See the last page for full reference in Russian.
Special thanks to Professor Anatoliy Maslak.

In many ways, educational assessment is still being done the way that it was done 50 years ago. Many organizations provide linear test forms of multiple-choice items or other traditional formats and score them with classical methods. While the advent of the computer has most definitely affected how tests are delivered, in many cases, it is still a traditional test with classical scoring, just shown on a computer screen rather than paper. With the coming of The Cloud, innovation is becoming stronger and faster, offering us more avenues to improving student assessment – and therefore student learning. In fact, there are three conferences in late 2017 devoted specifically to this topic: the International Association of Computerized Adaptive Testing, the MARCES conference at the University of Maryland, and the ACTNEXT symposium on computational psychometrics.

The rapid increase in computing power and availability of other technological resources is fueling this change, impacting nearly every aspect of assessment from the writing of the first item to the reporting of the final student score. This chapter will discuss a number of these aspects and some of the innovations currently under way, including automated item generation, automated test assembly, computerized adaptive testing, and psychometric forensics.

Item Development

Item development is arguably the most time-intensive portion of the test development cycle. High quality items can be quite expensive, so any improvement in the process that can reduce the effort is likely to result in a positive return on investment for the test sponsor.

The most straightforward innovation in this phase is a strong item banking platform; that is, one which makes it easy to author items, review them, leave comments, and store metadata. While these are traditional aspects of item development, a well-designed software system can massively streamline the process, contributing to both a cost decrease and a quality increase.

A more innovative technology is *automated item generation* (Gierl & Haladyna, 2012), which utilizes algorithms to build items. These use what some researchers call *item skeletons* or *item templates* to split an item into static and dynamic components. Figure 1 shows such a template from Gierl, Lai, Hogan, and Matovinovic (2015). Elements that are in bold/brackets are considered dynamic, and the item writer can then specify possible values.

Figure 1: Example Item Template

<Name>is coloring a **<Product.Name>**using **<Product.Material>**shared with **<Gender>** friends. Each of **<Gender>**friendshasthe same number of **<Product.Material>, <Product.number>.**There were **<Product. Material>**left over after **<Name>**handed them out to **<Gender>** friends.Which of the following equations represents this situation?

The simplest implementation of this innovation is to apply it to the creation of a number of fixed items in the bank. For example, the template above could easily be used to create 10 different items which are then stored in the bank, with some notation that they are variants (and therefore enemies which cannot be used on the same form), and added into linear test forms as

needed. Because the additional cost of the variants over the initial template is relatively small, the overall cost per item in the bank is dramatically decreased while quality remains constant.

A much more powerful implementation is that the test delivery platform be designed to create such items on the fly. This massively increases the number of possible variations (see the article for examples), which in turn massively increases the size of the item bank and the security of the assessment. Of course, the technology needs of the platform are therefore quite substantial, and very few testing platforms support such technology.

New Assessment Formats

At the item level, another important innovation is in item types and formats. The initial development here was to move from multiple choice items to newer but still fixed-format items such as multiple response. Later innovations included automatically scored open response items, drag and drop items, and hotspot items. More recent innovations are driving towards simulations, gamification, and performance testing.

An early example was the Joint Commission on Allied Health Personnel in Ophthalmology, which implemented simulated performance tests before 2006; these used a video-game like experience to mimic the use of ophthalmic instruments such as a phoropter or lensometer, letting the candidate turn dials and flip switches with their mouse while built-in artificial intelligence simulated the reactions of the patient or objects. An assessment like this greatly standardizes the experience, which is an important aspect of validity. It can also provide substantial savings if the alternative is to fly candidates to a real clinic with live patients on which to test.

Automated Test Assembly

One of the most manual tasks for a psychometrician is form assembly. Assembling one form is easy, but if the requirement is eight forms, parallel in both content and statistics (classical or IRT), and overlapping by 20-30% each but "spiraled" rather than a fixed anchor item block? This is exactly the sort of work for which we have computers. Some organizations rely on simple, straightforward algorithms, but those with a greater number of constraints need to leverage integer programming methods from the field of operations research.

Test Scheduling

Test scheduling and related operations like payment processing has historically also been a laborious task, though not for psychometricians. Instead, an organization needed to employ a call center full of people. Now, software platforms exist to automate much of this, allowing you to define things like test availability, prerequisites, pricing, locations, and proctors. Moreover, the transition to the API economy (application programming interface) allows related systems to cover some or all of these topics and then connect to the assessment platform. For example, a school district stores its primary information in a student information system (SIS) which could then automatically connect to an assessment platform to schedule. A similar situation exists for

applicant tracking systems in pre-employment testing and association management systems in professional credentialing.

Test delivery

Test delivery is unique in that it is the portion of assessment platforms that handles the most touches. That is, an organization might have a few or even a few dozen item writers, but likely has hundreds to millions of test-takers, and in many cases these are taking multiple tests throughout the year. This situation makes it the most ripe for improvements in innovation that allow for improved precision and other beneficial aspects of assessment.

Technology-enhanced items, as previously discussed, are one such approach. TEIs are typically intended to improve assessment by targeting higher order thinking or other constructs not easily assessable by simple item types. They are also often designed to increase examinee engagement and face validity. However, this does not necessarily mean that they deliver more bang for their buck in terms of measurement precision. Perhaps the same could be said about widgets like protractors and rulers; these simply emulate paper-based exams from decades ago.

Another avenue to improve assessment is the use of algorithmic test delivery such as computerized adaptive testing (CAT; Weiss & Kingsbury, 1984) or linear on the fly testing (LOFT; Becker & Bergstrom, 2013). Such approaches publish the test as an algorithm that creates a unique set of items for each examinee based on psychometric and content constraints. CAT does so in an interactive way, by utilizing item responses so far in the test to decide which item will be administered next. A related approach called multistage testing also does so, but adapts in blocks of items rather than after each individual item. Such tests can even be variable-length; some examinees might finish after 50 items, some after 100; it depends on the psychometric parameters. LOFT utilizes similar psychometric paradigms, but constructs a linear test form –a fixed set of items – for each examinee as the test begins. This way, each examinee will receive a different set of items, for example a set of 100 items from a pool of 300, but each set will be equivalent from a psychometric and content perspective. This does not provide the test shortening advantages of CAT but can vastly increase the security of the testing program.

Test proctoring

Security is also highly dependent on proctoring and other protocols that surround the test. Technology is also being leveraged here, often in an arms race against the technology used by examinees in cheating and other avenues of test fraud.

Perhaps the most prevalent innovation is the use of webcams as a conduit for virtual proctoring. While widely accepted for low and medium stakes exams, it is not yet fully accepted for high stakes exams, as there is often no way to fully control the test taking environment as the examinee is typically in a place of their choosing, like a home office or a library. Technology is also present in security features like a lockdown browser, IP address limits, and time limitations (full test, per section, or per item).

Psychometric forensics

While the cliché of “an ounce of prevention is worth a pound of cure” is well known in test security, it is because it is quite true. And while deterrent measures like those above are

essential, they do not prevent all test fraud. Psychometric forensics is often useful as a back-end approach for evaluating the possibility of test fraud.

Psychometric forensics is a family of quantitative analyses that are designed to detect examinees that are providing invalid responses. In many cases, this is innocuous, like low motivation, but in many cases it is cheating at the individual or even group level. While extant in the scientific literature for at least the past 50 year, it is only recently becoming more widely applied due to the availability of relevant software like SIFT (Thompson, 2016) or CopyDetect (Zopluogu, 2012). The future will see more automation, such as the availability of APIs or real-time dashboards, perhaps integrated with other aspects like virtual proctoring.

Essay scoring

Another extremely laborious stage in the test development cycle is that of scoring open response questions, which are typically essays. Some organizations still utilize the same processes they did 50 year ago, with stacks of papers being routed through a small army of readers sitting at folding tables in a large room. Such a process can take weeks or months. Technology can immediately make this process more efficient by routing images or html text to the readers and sending their ratings and comments directly back to a central database for scoring an analysis. However, the true opportunity for innovation is in automated essay scoring (Shermis & Burstein, 2003). This field saw a jump in innovation with an international competition sponsored by the Hewitt Foundation in 2012.

How test scores are used

While not always considered part of a psychometricians's domain, the actual use of test scores is nevertheless part of the validity context. It is also ripe for innovation. Many of the buzzwords we hear in the more general media – big data, machine learning, data science – actually describe what has been done to test scores for decades. The massive innovation occurring in those more generalized spaces is easily applicable to psychometrics. From small packages in R and Python to Google's TensorFlow system, the readily available software is changing quickly. This will affect how test scores are used to predict things like job performance, university graduation, and school evaluations.

Summary and Conclusions

As is evident from the brief survey of topics above, there is extensive innovation that is occurring in assessment, especially from a technological perspective. Nevertheless, the testing industry is a relatively conservative field, such that even if innovations are developed in academia or technology becomes widely available, the improvements are not always disseminated in a way that improves assessments for the millions of people that take tests every day.

What are some of the hurdles that face us, providing potential roadblocks for the implementation of innovation to improve assessments for examinees? The largest hurdle is availability. For example, automated essay scoring has been used by large organizations for at least 20 years, and adaptive testing has been used by large organizations for at least 35 years, but both remain relatively inaccessible to most organizations. Why? Perhaps the most

prominent reason is the sophistication of such approaches and therefore the lack of practitioners with sufficient expertise. The lack of software to apply innovative methodologies in digestible interfaces, for both input and output, is also a cornerstone of the problem. More and better software will reduce the barriers to entry, by lowering the level of expertise needed to implement certain approaches.

Yet this is an exciting time, both because of the opportunity that this problem presents as well as the vast amount of actual innovation that is happening. The most important warning is that we keep the end goal in mind, which is to improve the precision/reliability of scores and the validity of their interpretations. There have certainly been a number of innovations in the field that were driven by pedagogy and the technology itself, without that goal in mind. A prime example is some of the tech-enhanced item types that have been suggested, which in some cases actually violate the assumptions of psychometric models. While some of the innovations discussed are targeted more towards the bottom line of the testing organization, let us not forget the root directive.

References

Becker, K., & Bergstrom, B. (2013). Test administration models. *Practical Assessment, Research & Evaluation, 18*(14). Available online: http://pareonline.net/getvn.asp?v=18&n=14

Gierl, M.J., & Haladyna, T.M. (2012). Automatic item generation: Theory and practice. London: Routledge.

Gierl, M.J., Lai, H., Hogan, J.B., & Matovinovic, D (2015). A method for generating educational test items that are aligned to the Common Core State Standards. *Journal of Applied Testing Technology, 16(1).* http://www.jattjournal.com/index.php/atp/article/view/80234

Shermis, M.D., & Burstein, J.C. (2003a). Introduction. In M.D. Shermis & J.C. Burstein (Eds.), Automatic essay scoring: A cross-disciplinary perspective (pp. xiii-xvi). Mahwah, NJ: Lawrence Erlbaum Associates.

Thompson, N.A. (2016). User's manual for SIFT: Software for investigating test fraud. Minneapolis: Assessment Systems Corporation.

Weiss, D. J. & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.

Zopluogu, C. (2012). CopyDetect: An R package for computing statistical indices to detect answer copying on multiple-choice examinations. *Applied Psychological Measurement, 37*(1), 93-95.

ASSESSMENT SYSTEMS
FOR GOOD MEASURE™

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФИЛИАЛ КУБАНСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА В Г. СЛАВЯНСКЕ-НА-КУБАНИ
ЛАБОРАТОРИЯ ОБЪЕКТИВНЫХ ИЗМЕРЕНИЙ

ТЕОРИЯ И ПРАКТИКА ИЗМЕРЕНИЯ И МОНИТОРИНГА КОМПЕТЕНЦИЙ И ДРУГИХ ЛАТЕНТНЫХ ПЕРЕМЕННЫХ В ОБРАЗОВАНИИ

Материалы XXV Всероссийской (с международным участием) научно-практической конференции (г. Славянск-на-Кубани, 09–10 июня 2017 года)

Славянск-на-Кубани
Филиал Кубанского государственного университета
в г. Славянске-на-Кубани
2017