Policy Brief
October 2012

United Nations Educational, Scientific and Cultural Organization
UNESCO Institute for Information Technologies in Education

UNESCO Institute for Information Technologies in Education

HOW TECHNOLOGY CAN CHANGE ASSESSMENT

CONTENTS:

STATEMENT AND SCOPE OF PROBLEM
USES OF ASSESSMENT
ASSESSING LEARNING EXPERIENCES VS. ASSESSING INDIVIDUALS
FORMATIVE VS. SUMMATIVE ASSESSMENT
WHAT CAN BE ASSESSED AND HOW?
CHALLENGES AND FORWARD PROGRESS
RECOMMENDATIONS AND CONCLUSIONS

STATEMENT AND SCOPE OF PROBLEM

Assessment is a large topic for Information and Communication Technology (ICT), with important issues ranging from how to evaluate the effectiveness of an ICT to how to ensure that someone receiving certification through an online course has proven mastery. One approach to assessment with ICTs is to use technology to make existing forms of assessment more reliable and efficient. This brief concentrates instead on how technology can change the way we think about assessment, addressing two specific issues. First, how can technology more effectively enable assessments to be used as a tool to improve student learning in ICTs? Second, how can technology increase the range of learning outcomes we can assess and what will these assessments look like?

USES OF ASSESSMENT

Assessment is a big enterprise involving many possible stakeholders. For example, the Programme for International Student Assessment (PISA) is an assessment given to students in over 70 countries to provide a common metric of student performance. It yields results at the national level. In contrast, teachers use informal formative assessments in their classrooms every day to gauge individual student progress and tailor their instruction appropriately. Both of these assessments, as well as others, have value for different purposes and stakeholders. It is important to understand the goals and stakeholders of an assessment to make appropriate decisions about its design, the intended use, and the interpretation of the data it produces.

Below are a few examples of the potential functions of assessment and related stakeholders:

PURPOSE	EXAMPLE	PRIMARY STAKEHOLDERS
Ranking a country (state, district) against others	The Programme for International Student Assessment (PISA)	Governments, politicians, policy makers
Measuring progress of a school (district, state) over time	API growth indices in California state standardized tests	Governments, politicians, policy makers, school administrators
Ranking individuals	IQ tests, the SATs	Students, employers, admissions officers
Measuring an individual's mastery of a subject matter or skill set	Certification exams; IB tests	Students, teachers, degree/certificate granting institutions, schools, employers
Understanding a student's current skill and knowledge state to more effectively tailor instruction to that student	Embedded formative assessment in adaptive tutoring systems	Students, teachers, instructional designers
Measuring the effectiveness of an instructional intervention or learning environment	Randomized controlled trials, such as in the What Works Clearinghouse http://ies.ed.gov/ncee/wwc/	Policy makers, instructional designers/ instructors
Iterative improvements to an instructional intervention or learning environment	A/B testing, in which students receive different versions of an intervention to test which is more effective	Teachers, instructional designers, software companies

This brief discusses innovations in assessment made possible by technology that can apply to all the levels of assessment described above. However, in relation to ICTs specifically, the bottom entry in the table – the use of assessment for iterative improvement – is underutilized. We highlight this new possibility as a major way to transform the design, evaluation, and delivery of ICTs.

ASSESSING LEARNING EXPERIENCES VERSUS ASSESSING INDIVIDUALS

Discussions of assessment in ICTs often concentrate on increasing the precision of assessments for evaluating individuals. One reason is that test makers are concerned about issues of reliability (consistency) and validity (are they measuring what they intend) of online assessments, and also about how to prevent cheating. These are important issues. For example, as online programs increasingly confer certification, it is crucial that institutions can be confident that the students have actually mastered the material. A second reason for improving the precision of measuring individual learning is that it enables adaptive learning systems. Adaptive learning systems ensure on-going assessments of a learner's knowledge state, so they can tailor problems and instruction appropriately. Cognitive tutors (e.g., http://pact.cs.cmu.edu), for instance, continually assess students, updating a model of the student's skills and knowledge. This model of the learner is compared to a model of an expert, so the system can determine discrepancies and deliver subsequent assignments accordingly.

A different goal of assessment is to evaluate learning experiences. Instead of making an inference about how well a student is doing, the desired conclusion is about how well instruction is doing. This is not a new idea; educational research has been comparing different instructional conditions for decades. However, when one explicitly shifts the focus of assessment to learning experiences rather than individuals, there is a shift in priorities. Questions about the reliability and validity of assessment measures remain important, but equally important is the ability to make inferences about the quality of different specific learning experiences. If the goal of assessment is to improve instruction, then the measures need to yield actionable information at the level of the learning environment.

FORMATIVE VS. SUMMATIVE ASSESSMENT

Educational researchers and policy makers have recommended increased formative assessment in classrooms and learning environments (e.g., Darling-Hammond, 2010; US Department of Education, 2010). Often considered assessment "for" learning (as opposed to assessment "of" learning), the goal of formative assessment is to provide feedback during the learning process to inform instructional decision making. For example, formative assessment might help a teacher realize which part of a math formula students are struggling with during the lesson, so the teacher can shape instruction accordingly. In adaptive learning environments, formative assessment allows the environment to gauge the student's current knowledge state to determine what problems and instruction to present next. This is in contrast to summative assessment, where the goal is to provide information about the end product of learning – did the student learn to a standard or not? Standardized tests that report the performance of students in a given school district are an example of summative assessments. These tests are evaluative, and there is no assumption that information from them will inform the instructional process for a particular student.

Shifting the assessment focus from individuals to learning environments has implications for the design of research. Large-scale randomized controlled trials, in which students are randomly assigned to treatments to evaluate the effectiveness of an educational technology, are generally summative assessments. The goal of the assessment is not to improve the learning environment; rather, the goal is to determine whether the learning environment increased scholastic gains compared to no instruction or a different form of instruction. The treatment comparisons for these types of studies are usually too coarse to determine which aspect of an instructional treatment made a difference one way or another. As such, they cannot serve a formative function, but rather, they are designed to decide whether it makes sense to adopt a particular technology or curriculum.

Technology makes it possible to use large-scale testing in a more formative way to help shape and improve the effectiveness of ICTs for learning. In particular, the ability of ICTs to reach a broad range of students, collect data, and present different variants of the same material makes for a powerful research tool. Questions as broad as the best trajectory through the learning content or as narrow as the most effective way of visualizing information can be tested empirically, and the environment can be shaped to reflect the results. Under this formative model of research, students can still be randomly assigned to conditions, but the goal of assessment is the continual improvement of the ICT, not the final proof of effectiveness.

One way ICTs enable large-scale formative research is through the use of A/B testing. A/B testing is used in the software industry to determine optimal features for products and marketing. For example, Amazon.com might show a million customers a webpage for a given product that it is selling. For half of the customers the webpage might show a red button, and for the other half of the customers there could be a green button. If more customers who see the green button buy the item, Amazon has learned that the green button configuration is more effective. These ideas can be applied to education in addition to marketing. Refraction (http://games.cs.washington.edu/Refraction) is a game-based approach to designing an optimal sequence of problems for fraction learning. By using A/B testing, the Refraction system can determine the optimal sequencing of the curriculum. For example, after completing problem set A, half of the students can be sent to problem set 1, while the other half works on problem set 2. Afterwards, all the students can complete problem set B. One can then compare which sequence leads to better performance on problem set B: A1 or A2. If the students do better after A1, this sequence can then be incorporated into the design of the system.

As a second example of iterative design and testing, the Learn Lab at the Pittsburg Science of Learning Center (http://learnlab.org/) used A/B testing to modify and improve an online chemistry class over the course of a few semesters. Students received versions of the course that differed in the presentation of diagrams. Embedded learning assessments, patterns of usage time and qualitative interviews were analyzed. The results of these comparisons then fed into further course redesign and additional testing (see, for example, http://evidenceframework.org/rapid-controlled-experimentation-to-facilitate-learning-difficult-conceptual-content).

While many decisions in the design of ICTs can be informed by existing data and learning theory, there are countless decisions to be made that will influence learning, and for which no empirical research currently exists (or the research is conflicting). Trying to run studies to address all of these decisions in advance of developing a learning environment is intractable. A different approach involves an iterative process of controlled experimentation and continual improvement that is based on data collected during the use of the ICTs. This type of bottom-up design research requires very large samples of students to test the many possible combinations. Technology, especially the internet, helps to solve this problem. Crowd-sourcing across thousands of internet users has rapidly become a standard tool within the research kit of behavioral scientists. There are even commercial-grade platforms, such as Amazon's Mechanical Turk (www.mturk.com), that help researchers connect with willing users to conduct A/B experiments.

WHAT CAN BE ASSESSED AND HOW?

Technology has the power to transform what is assessed in ICTs and how. Assessments can go beyond end-of-unit multiple-choice questions to gather process data on how students are going about learning and solving problems. This occurs by embedding learning assessments within the learning experience and analyzing process data in log files that capture every click and key stroke. In reference to computer learning games, Shute (2011) uses the term "stealth assessments," to refer to embedded assessments that collect user performance data as part of the game. The performance data enables inferences about the user's knowledge state and learning. This kind of assessment can reduce test anxiety, because the lines between learning and assessment are blurred; the assessment is integrated with the learning process. It also has a benefit for overall time spent learning instead of testing, because it is not necessary "to close the store to take inventory." For example, one task in a game about river ecology involved students collecting virtual water samples from various parts of a river. The system logged the accuracy and efficiency with which students collected the samples. The system further used these data to inform the prediction of a student's information gathering abilities.

Embedded assessments can help make ICTs adaptive to individual student performance and skill levels by allowing dynamic modification of instruction and difficulty based on performance. They can also inform the design of learning experiences by comparing the outcomes of the embedded assessments for students who received different versions of the environment. By collecting data that indicates accuracy, hints needed, time on task, resources used, and so forth, embedded assessments can provide designers with evidence of the effectiveness of design choices for learning and engagement, while minimizing disruption and test anxiety.

It is important to note that embedded assessments do not need to be hidden assessments. In fact, there are examples where providing students with the results of embedded assessments can drive greater learning and engagement. For example, the popular online game World of Warcraft continually assesses player progress and presents feedback to the player in the form of heads up display that appears on the game screen. The information is highly motivating and points students to where they should focus their attention and learning efforts so they can do better and open up new levels within the game (Reeves & Read, 2009).

In addition to reconsidering when and where assessment can happen, we can also rethink what is being assessed. While codified knowledge is important, equally crucial

for 21st century job success is the ability to adapt and learn (e.g., Benner, 2002). Bransford and Schwartz (1999) introduced the concept of preparation for future learning assessments. In this kind of assessment, rather than only measuring the student's codified knowledge (e.g., can they come up with the formula for momentum), students are provided learning resources during the assessment. The assessment measures how prepared students are to learn from those resources. For instance, they might be given a worked example of a momentum problem within the assessment, and then asked to solve a new kind of problem they had not yet encountered, but for which learning from the worked example would be helpful. For iterative design of ICTs, the question becomes whether one experience (i.e., one version of the system) prepared students to learn new, related content better than another. These kinds of assessment measures are more similar to what students will experience outside of school (e.g., learning on the job) and can show differences between learning experiences where traditional measures of codified knowledge do not (e.g., Chin et al., 2010).

Technology can allow researchers and designers to collect data on students' choices about learning, which creates an interactive preparation for future learning assessment. For example, in a simulation environment called River City, researchers used data about what kind of information students chose to examine within the simulation to assess their inquiry skills (e.g., Ketelhut, 2007). Schwartz and colleagues (2012) used computer log files to examine how students used critical thinking skills to learn. Students needed to mix the primary light colors – red, green, blue – to create new colors. The online environment provided several resources for learning including an experiment room where students could mix light beams as well as a set of charts that held conflicting information for how to mix light to create different colors. Students who spent more time evaluating the charts learned more about mixing light and were also doing better in school.

In the United States, the National Assessment of Educational Progress (NAEP) is a large nation-wide assessment of student performance. NAEP is piloting assessments in which they track students' use of resources, such as charts and graphs, during science inquiry tasks (e.g., what causes phytoplankton growth). Measuring learning process data can provide a powerful indicator of a given students' inquiry and critical thinking skills. Across students, collecting data about learning choices and outcomes makes it possible to examine which learning patterns are more productive (lead to better learning outcomes). This information can then be used to inform design choices within the ICT. For example, the system can intervene if students appear to be heading down an unproductive learning path.

CHALLENGES AND FORWARD PROGRESS

An important challenge in collecting process data from ICTs is determining how to analyze it. One solution receiving considerable attention is data mining. Data mining is an approach that leverages computers to discover patterns in large sets of data. Data mining is used in business, for example, to find patterns in shopper's buying habits. In educational applications, an example might involve using click-stream data from computer logs of student interactions along with information about the students' prior achievement to determine the best kind of student help to provide within an online system. For example, using student data from interactions with an intelligent computerized tutor, Min Chi and colleagues conducted data mining and machine learning to infer and design teaching policies about which action the computer tutor should take for a given problem-solving event to optimize student learning gains (e.g., tell the answer to a problem or ask the student to answer, Chi, VanLehn, Litman, & Jordan, 2010). While data mining is becoming a hot topic in education, there are still considerable impediments to its wide scale adoption. The required computational and human resources are still expensive. The process requires a considerable amount of human intuition, knowledge, and effort on the front end to determine the most effective way to structure the information that goes into the machine learning algorithms so they will generate meaningful and actionable patterns. As data mining techniques advance in business and education, they will likely become increasingly available and tractable for wide-scale application in ICT assessment.

One example of an effort to reduce the expense and increase the tractability of assessments of complex phenomenon comes from the autoscoring of writing samples and the crowd-sourcing of answers. For example, the online course system, *Coursera* (https://www.coursera.org), can register more than 10,000 students in a class. They have started using peer grading and crowd-sourcing to make grading of writing samples possible. Students in the class are trained to use a grading rubric, and are assigned to grade each other's work. Multiple students are assigned to grade each work. This is based on crowd-sourcing research, which found that taking the average of multiple ratings will lead to an accurate score, even if the accuracy of any individual rating may not be reliable. As another approach, the educational testing service (ETS) is automating the grading of writing and open-ended responses using natural language processing (http://www.ets.org/research/topics/as_nlp). Natural language processing refers to using computer algorithms to make sense of and interact with human language. This is being applied to students' written responses to analyze the content in relation to a rubric or intended concept.

Tractability and the ability to scale-up are not the only challenges to the new wave of technology-enabled assessments. With increases in the ability to seamlessly measure learning with embedded assessments, there are new ethical concerns about the fairness of the tests. Students may not be aware they are being evaluated or what aspect of their performance is being evaluated, and therefore, they may not show *maximal* performance (what they can do under optimal circumstances when they know they are being assessed). This is an important concern when the goal of assessment is to sort and characterize students, for example, to determine who gets into a university program or enrichment program. Students may not put forth their maximum effort, because they do not know that there is a high-stakes assessment occurring. As a result, the assessment may not

provide a good characterization of student potential or accomplishment. However, if the goal of an assessment is to inform the design of instruction rather than rank students, then the fairness concerns are less pressing. For the goal of improving instruction, it may be more effective to collect data on students' *typical* performance (what they would be likely to do in a new situation in absence of explicit assessment cues). On the assumption that instruction occurs under conditions of typical performance, then assessments of typical performance would be the most informative.

Finally, when using assessment for continual improvement of ICTs as proposed in this brief, it is important to avoid hurting students by using instructional conditions that are known to be ineffective or put student learning at risk. For instance, with A/B testing, it is important to avoid design decisions known from research to lead to less effective learning. However, for many design choices, there is no research available about which is better, or the research is incomplete or conflicting. In these cases, testing different versions will be unlikely to hurt students relative to current practice, because students will often receive "standard practice" as the alternative condition. Moreover, given the results of a comparison, the more effective practice can be rapidly incorporated into the learning environment to help future learners. For these kinds of decisions for which there is no directly applicable research, rather than making a command decision and sticking to it, as is often the practice, empirically testing different options leads to ultimate improvement of the learning environment.

RECOMMENDATIONS AND CONCLUSIONS

Technology can increase the efficiency of existing models of assessment, but it can also transform how assessment is used and conceptualized. This brief makes the following recommendations:

- Assessments should be designed to yield information that is actionable at the appropriate level. Assessments have many different uses, from ranking nations to characterizing a student to evaluating a learning experience. Different uses require different kinds of assessment. For example, assessments designed to rank schools do not generally provide information that is useful at the level of making instructional decisions for an individual student. One frequently overlooked use of assessment is for the purpose of informing the design of learning environments. In this case, assessment should be of a sufficiently fine granularity to isolate the factors in the learning environment that lead to increased or decreased learning.
- Formative assessment can be an important tool for making instructional decisions. Formative assessment is considered assessment "for learning," where the primary goal is to inform the learning process. This contrasts with summative assessment, which measures the end product of learning, and is primarily evaluative. Formative assessment can be used at an individual level to decide what information or problems to present to a student given their current level of understanding. It can also be used at the level of ICT development to determine which design choices are most effective for learning in a virtuous cycle of testing and improvement.

- Assessments can be embedded in ICTs. They do not have to take time away from learning. The embedded assessments can be integrated into the learning process, such that students are not aware they are being assessed, which can reduce test anxiety. Or they can be explicit, such that students are aware they are being assessed and can see their progress and improvement.
- One way ICTs can leverage assessments to inform the continual improvement and refinement of the ICT is by taking advantage of the internet, using A/B testing and crowd-sourcing to gather feedback and improve the design of the learning environment. Varying features of an ICT in a controlled way allows for evaluation of which design and instructional choices lead to the best learning.
- In addition to collecting traditional measures of student knowledge, ICTs can measure process data about how students go about learning and solving problems in the ICT. This allows for assessment of student inquiry skills and students' preparation to continue learning outside of the ICT, which is in line with the 21st century demand to keep learning on the job. Looking at student process data can also inform which patterns of interactions lead to better learning outcomes within the ICT, which can help ICT developers design productive sequences of instruction. Investment in new methods of data mining will make it possible to analyze and evaluate important process outcomes, such as how students go about learning, expanding what is possible to assess.

SUMMARY

One approach to bridging technology and assessment involves using technology to make existing models of assessment more robust and efficient. Another approach is to use technology to change how we think about the function and reach of assessment. This brief takes the later approach. Technology-aided assessments can help in the design, delivery, and improvement of learning environments. A common model of ICT assessment involves large-scale randomized controlled trials, in which a large sample of students is randomly assigned to a treatment and control group to test the effectiveness of an educational technology. This type of summative assessment may be useful for determining whether to spend millions on a large-scale adoption of an existing technology, but it is not an efficient way to help in the development of new innovative technologies that can shape the future of education. ICTs have a unique ability to reach many learners (for example, *Coursera* has over 1.5 million users). Additionally, they can be designed to collect data and selectively present different information to different students. This can allow for the empirical testing of questions about the best design choices for learning, and the results can be incorporated into the learning environment for continual improvement. In terms of reach, technology allows for new kinds of assessments, including assessments of student's learning and problem solving processes that are embedded in the learning context, and assessments of how well prepared students are for future learning. Assessment has a powerful effect on education, not only serving an evaluative function, but also shaping what is considered important for students to know. Expanding the scope of assessments from evaluating only end-state knowledge to evaluating learning processes themselves has the potential to transform what is taught and how.

REFERENCES:

Benner, C. (2002). Work in the new economy: Flexible labor markets in Silicon Valley. Oxford: Blackwell Publishing.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad & P. D. Pearson (Eds.), Review of Research in Education, 24, 61-101. Washington DC: American Educational Research Association.

Chi, M. VanLehn, K. Litman, D., and Jordan, P. (2010). Inducing Effective Pedagogical Strategies Using Learning Context Features. Proceedings Eighteenth International Conference on User Modeling, Adaptation, and Personalization (UMAP2010). Pp. 147-158.

Chin, D. B., Dohamen, I., Oppezzo, M., Cheng, B., Chase, C., & Schwartz, D. L. (2010). Preparation for future learning with Teachable Agents. Educational Technology Research and Design, 58, 649-669.

Darling-Hammond, L. (2010). Performance counts: Assessment systems that support high-quality learning. Washington, DC: Council of Chief State School Officers.

Ketelhut, D. J., (2007). The impact of student self-efficacy on scientific inquiry skills: An exploratory investigation in River City, a multi-user virtual environment. The Journal of Science Education and Technology, 16 (1), 99-111.

Reeves, B. & Read, J. L. (2009). Total engagement: Using games and virtual environments to change the way people work and businesses compete. Cambridge, MA: Harvard Business Press.

Schwartz, D. L., & Arena, D. (in press). Measuring what matters most: Choice-based assessments for the digital age. Cambridge, MA: MIT Press.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503-524). Charlotte, NC: Information Age Publishers.

U.S. Department of Education. 2010a. National Education Technology Plan. http://www.ed.gov/technology/netp-2010.

Many discussions of technology-based assessments concentrate on automating current methods of testing to save time and cost. However, technology also changes what educators can assess, how and when to assess it, and for what purpose. Assessments can be embedded in ICTs, and have the potential to measure learning processes, in addition to end-of-lesson knowledge. Technology-aided assessments are useful not only in the evaluation of ICTs, but also as part of the design process, leading to iterative improvement. This brief focuses on assessment in ICTs, discussing how technology-enabled assessments can be leveraged to improve ICT design and student learning.

Authors: Kristen Blair and Daniel Schwartz

Published by the UNESCO Institute
for Information Technologies in Education
8 Kedrova St., Bldg. 3
Moscow, 117292
Russian Federation
Tel: +7 (499) 129 29 90
Fax: +7 (499) 129 12 25
E-mail: Liste.info.iite@unesco.org
http://www. iite.unesco.org

Printed in the Russian Federation

ISSN 2221-8378