E–Learning and Digital Media
Volume 8 Number 2 2011
www.wwwords.co.uk/ELEA

The Nature of Assessment Systems to Support Effective Use of Evidence through Technology

KATHLEEN SCALISE
University of Oregon, USA
MARK WILSON
University of California, Berkeley, USA

ABSTRACT The National Educational Technology Plan 2010 (NETP) presents a model of twenty-first-century learning powered by technology, with goals and recommendations in five essential areas: learning, assessment, teaching, infrastructure, and productivity. This article connects NETP ideas in one of these areas – assessment – with those of the Assessment and Teaching of 21st Century Skills (ATC21S) project. Launched by Cisco, Intel and Microsoft, ATC21S released results of a methodological working-group study at the Learning and Technology World Forum 2010 in London. The ATC21S methodological report discusses how the development of a good assessment system is rooted in the inferences the system is intended to support. Assessment is a special kind of evidentiary reasoning with evidence used to support particular kinds of claims. This article will illustrate how key questions and answers for decision-making are influenced by a new era of educational assessment with technology. Areas discussed include the characterization of the constructs to be assessed, the kinds of instruments to be developed, the level of information gathered, and promising avenues for analytic approaches.

The National Educational Technology Plan, 'Transforming American Education: Learning Powered by Technology', includes a number of goals and recommendations for the future of US educational assessment. It stresses providing timely and actionable feedback about student learning in the classroom. Data are to be used directly to improve achievement and instructional practices for individual students, rather than primarily as a post-intervention accountability process. The intent is to serve not only students and teachers but also a variety of other educational stakeholders, such as school administrators and the state, for continuous improvement. Goals of the plan emphasize the need for much richer and more complex tasks, building capacity for use and research on assessments in the schools, and revising policies and regulations to protect privacy while enabling effective data collection and use.

These are laudable goals. They are also challenging. When considering educational assessment, perhaps one of the most important yet most commonly overlooked components is the issue of how to present results to various types of stakeholders. This is of prime importance, as it is from the results of an assessment that decisions are made which influence the future knowledge acquisition of test takers. Reflecting on the kinds of reports on assessment results that we aspire to provide can present us with an excellent starting point to think about the challenges that we must face in the design of assessment structures to support the development of twenty-first-century goals such as described by the NETP.

In this article we look at the NETP goals and recommendations through the lens of the Assessment and Teaching of 21st Century Skills (ATC21S) project. Launched by Cisco, Intel and Microsoft, ATC21S released results of working-group studies at the Learning and Technology World Forum 2010 in London (Wilson et al, 2010a). The ATC21S methodological report discusses

 http://dx.doi.org/10.2304/elea.2011.8.2.121

how the development of a good system for skills involving higher-order thinking is rooted in the inferences that the system is intended to support. If we consider what results need to be reported to serve the needs of the NETP goals described above, this can facilitate progress in meeting challenges and achieving the intended results.

The Twenty-First-Century Constructs

There have been many efforts to create lists of skills that have been taking the label of twenty-first century. In the United States, these have been framed about career and college readiness. The ATC21S project provides a perspective on a range of twenty-first-century skills, based on a comparative study of the research literature in this area. Some examples of typical twenty-first-century skills from the project that students need to know and be able to do include collaboration (teamwork), creativity and innovation, and information literacy (Binkley et al, 2010). For the purposes of the NETP, twenty-first-century skills are most likely to be framed in terms of the Common Core State Standards Initiative. This is a state-led effort to provide standards to prepare students for college and the workforce. Coordinated by the National Governors Association Center for Best Practices (NGA Center) and the Council of Chief State School Officers (CCSSO), to date the adopted standards are in English language, arts and mathematics.

Why are the reports a good starting point? Because they encourage us to think about the topics that we want to assess, they invite us to consider what kind of uses and decisions (inferences) we want to promote among users, and they lead us to ponder what kind of evidence we deem appropriate to support those inferences.

The NETP acknowledges that for US schools today, most assessment data gathered for accountability are used after the fact, and not in the classroom during the learning process. Little accountability focus has been placed on assessing student thinking in the process of learning to help guide instruction.

The NETP also notes that rarely in the United States are data sets collected and aggregated 'in ways that make the information valuable to and accessible by educators, schools, districts, states, and the nation to support continuous improvement and innovation' (US Department of Education, 2010, p. 25).

To achieve these goals of classroom improvement, the NETP must focus on strong evidentiary reasoning (Mislevy et al, 2003, p. 25) as the starting point of a sound assessment. The NETP is a technology plan, and to bring about its goals, it must place emphasis on *measurement technology* as well as the traditional *information technology* (Wilson, 2003). Measurement technology includes: (a) defining the constructs that will be measured; (b) creating the tasks that will be used to elicit responses and performances; (c) interpreting or assigning values (codes, scores or other inferences) to the student responses to those tasks; (d) delivering and gathering the responses; and (e) modeling and analyzing those responses to generate the intended reporting to stakeholders. Such measurement technology is needed for valid, reliable, fair and feasible results for the intended uses. In addition, a goal of reporting in this context should also be to serve as broad a population of students as possible.

Changing thinking about what we should be assessing is a major priority of the NETP report. It calls for both standards and assessments to establish that US students 'possess 21st century skills like problem-solving and critical thinking and entrepreneurship and creativity' (US Department of Education, 2010, p. 26). Translating the domain of the Common Core State Standards into operational measurement constructs that address these skills is necessary, and will take some effort and research. Following the logic of measurement technology as described above means that the starting point for the assessment of twenty-first-century skills must be an adequate construct definition.

What is involved in an adequate construct definition for twenty-first-century skills? It goes beyond simply having standards. It is also necessary to show the relationships between and among the standards, and how mastery of the overall skill builds over time. In addition, it is necessary to describe which successful performances will be taken as evidence indicating the presence of the construct (to a certain extent), and which unsuccessful performances will be taken as evidence of the lack of the construct (to a certain extent).

Current literature in the field of educational assessment stresses that any measurement should be rooted in a robust cognitive theory. This includes a model of the learner that informs not only what counts as evidence of mastery but also what kind of tasks can be used to elicit this (National Research Council, 2001). A key aspect is the need for a developmental understanding of cognitive phenomena. This idea is laid out in the NRC report *Knowing What Students Know* (National Research Council, 2001): the term 'development' is critical to understanding the changes in children's conceptual growth. Cognitive changes do not result from mere accretion of information, but are due to processes involved in 'conceptual reorganization' (p. 112). The elaboration of definitions rooted in a conception of cognitive growth confers meaning on the ideas of 'improvement' and 'learning' when used in relation to the constructs. This is accomplished by describing and exemplifying what it means to become more proficient in each skill, serving as a base for the modeling of progress in each construct.

The NETP supports these ideas of drawing on a fuller understanding of human cognition in order to measure appropriately in these domains. It states that 'cognitive research and theory provide rich models and representations of how students understand and think about key concepts in the curriculum, and how the knowledge structures we want students to have by the time they reach college develop over time' (US Department of Education, 2010, p. 26).

While this acknowledges the importance of a thorough and deep understanding of constructs needed for fair and valid educational measurement in these complex areas, it also overstates how much progress is currently present in such cognitive theory for measurement. For the vast majority of areas that the NETP suggests for twenty-first-century assessment, the cognitive theory is lacking and the constructs have yet to be developed. The NETP does a good job of citing one important example from middle-school physics, in which one research team has done an extensive twenty years or more of qualitative work that has helped us understand how students learn, in this small portion of the standards (US Department of Education, 2010). Indeed, we have recently published research on how to improve reliability of the cognitive diagnoser based on this work (Scalise et al, 2010). However, it would be a mistake to believe this work exists broadly in other areas. The promise of an important future for this work exists but has yet to be realized.

Fortunately, there is a recommendation of the NETP report that calls for conducting research to explore necessary development for the NETP goals (Recommendation 2.3). This recommendation primarily describes information technology research needed. The point here is that *measurement technology* research, especially construct development, should also be an essential part of carrying out this recommendation.

It is worth noting that a major aim for the emphasis in the measurement research literature on cognitive development is to help teachers build a common conception of progress, serving as a base for the coordination of instructional practice and assessment. That may require a substantial shift in view for some instructors from a deficit model to a view of student learning as progressing toward richer and deeper understanding, with signposts along the way through the standards that help us understand how the learning process is proceeding.

When elaborating a developmental definition of a skill, the question remains about the characteristics that this kind of definition should have. What are the minimum elements that such a definition should address? A recent report from the Center on Continuous Instructional Improvement (CCII) about a specific kind of developmental perspective, *learning progressions*, presents a summary of characteristics that are desirable when defining a developmental model of proficiency (Corcoran et al, 2009):

- Learning targets: Briefly, describing what mastery is for a given skill is often the first step in the elaboration of a developmental definition.
- Progress variables: What patterns can be seen along the way as students begin to show what they know and can do, from emerging to moving towards mastery of the learning targets? How many and what kinds of different progressions can be seen? Are they similar or different for different learners?
- Levels of achievement: What types of scores or characteristics might be assigned to various stages of progress toward mastery? How can these interpreted as benchmark 'levels' of performance? What would they mean about being on target for learning goals?

- Learning performances: What examples provide operational definitions to specify the development of assessments? What activities would locate where students are in their progress?

Once the construct has been clarified, subsequent steps help allow the realization of NETP goals of assessment. These may be described in terms of the evidence-centered four-process architecture proposed by Almond et al (2002), and also exemplified by processes in the BEAR [1] assessment system (Wilson & Sloane, 2000; Wilson, 2005). Tasks are *selected and developed* on the basis of their relevance to the construct of interest, and *presented* to learners. Through engagement in the tasks, the learners generate evidence relevant to the *identified* construct of interest. Evidence from different sources (i.e. different tasks) is *accumulated*, which is then used to make inferences about the construct of interest. As described by Schum (1987), evidence here is data that increases or decreases the likelihood of the acceptance of a claim.

Twenty-First-Century Assessment Tasks

As Messick (1989) has suggested, a validity argument in assessment consists not only in showing that the evidence collected does support the intended inferences, but also in showing that plausible rival inferences are less warranted. This is where the specifications of the tasks are crucial, particularly in the context of twenty-first-century skills, which often must necessarily involve complex performances. The collection of tasks presented to students must be designed and assembled in such a way that *plausible rival interpretations* – such as the possibility that success was due to familiarity with the particular context rather than because of the underlying skill – are less warranted than the intended inferences.

Some of the twenty-first-century e-learning examples in the assessment chapter of the NETP may not employ sufficient measurement technology to adequately rule out these alternate hypotheses, especially if data are to be aggregated and used for continuous improvement planning in schools. Often in the e-learning context, the teaching and learning systems with embedded assessments were not designed for the purpose of aggregating for accountability, as suggested in the report. For instance, our work with Jim Minstrell on cognitive diagnosers found that some changes in the way the assessment content was designed in an example from that setting improved reliability substantially for one instrument studied (Scalise et al, 2010). The changes in design involved calibrating the difficulties of questions with quantitative measurement models, and then aligning the difficulty of the item pool with student abilities to better maximize item information and achieve a more reliable result. It will be important for the NETP vision that the technology products used in assessment employ measurement technologies sufficient for their intended use.

The NETP does correctly point out that many e-learning technology products exist that can begin to 'make visible sequences of actions taken by learners in simulated environments', and help 'to model complex reasoning tasks'. Making some actions visible is a good step along the way toward assessment but does not necessarily rise to the quality of evidence needed to justify fair, valid and reliable measurement. Technology provides a wonderful opportunity to improve our ability to use assessments in order to learn about students. NETP goals can help improve our capacity to elicit products or actions that will provide us with information about the constructs that interest us. The quality of the tasks that we use to stimulate this information about the progress variable is important, because it will determine whether we consider these observable responses as valid evidence of the proficiency level of the student.

The creation and selection of tasks plays an important role, not only for the obvious reason that they will ultimately constitute the assessment, but also because in many tasks, if not most, the construct itself will not be clearly defined until a reasonably large set of tasks has been developed and tried out with students. Simply stated, the design of the tasks helps clarify the construct that is being measured, bringing into focus ambiguities or aspects of the construct that have not been well described in the typical definition of a construct. This is not to diminish the importance of clear initial definition of the construct, but rather to recognize the role of evidence in the initial design phase in sharpening and, as necessary, reshaping the definition.

The relationship of the task to the construct is an important one. Typically, the task is but one of many that could be used to measure the construct. Where one wishes to represent a wide range of contexts in an instrument, it is better to have more tasks rather than less. This requirement has

to be balanced against the requirement to use item formats that are sufficiently complex to prompt responses that are rich enough to support the sorts of interpretations that the measurer wishes to make with the measures. And both requirements need to be satisfied within the time and cost limitations imposed on the measuring context.

As the NETP points out, assuming that we have (a) standards for the competencies students must have, and (b) sufficient cognitive knowledge to translate the domain of the standards into operational measurement constructs, then technology can be very helpful in assessing and rewarding learning regardless of when and where it takes place.

As described in the NETP, the evidence will include performance-based assessments. Following traditional paths of argument in assessment, this will lead to issues of cost, human scoring, inter-rater reliability, logistics of managing extensive work products, and so forth, as well as pursuit of a variety of artificial intelligence algorithms for improved automated scoring through the technology platform. These challenges have previously stemmed the influx of performance-based instruments into large-scale assessment. However, the NETP is correct in pointing out that 'the full flexibility and power of technology' is far from being tapped for designing, developing and validating new assessment materials (US Department of Education, 2010, p. 25). This is an exciting and promising area for educational measurement to make some substantial advances in.

New approaches to technology-mediated content, such as 'assessment objects', which are online learning objects specifically designed for evidence collection (Scalise & Gifford, 2008; Scalise, 2010), simulations, virtual worlds, sensors, and other virtual capabilities, expand what we might mean by performance-based opportunities for twenty-first-century contexts. Such approaches are definitely ready for more extensive research about their qualities that provide sound evidence for decision-making.

Entities such as the growing 'digital' divisions of the major educational publishing houses and other technology vendors are beginning to include online assessment opportunities embedded in their products that are being adopted by school districts as part of the standard curriculum process (Kulik, 2003; Robyler, 2006; Moore, 2007). All of these mean that there are many new opportunities for the measurement of complex constructs, and for the availability of large amounts of data, should planned sharing of data across contexts be enabled. New types of performances may suggest new acceptable routes to defining evidence, without incurring the same substantial barriers as previously was the case for exclusively paper-and-pencil or hands-on performance assessments.

The NETP does preserve a distinction between summative and formative assessment. Summative is defined as used 'to determine what students have learned for grading and accountability purposes'. Formative is defined as used 'to diagnose and modify the conditions of learning and instruction'. The measurement community should note that 'use' is established as the distinction, rather than quality of evidence, types of assessments, or provider. This is important because the NETP's strong goals for formative assessment to help drive instruction and to aggregate the data to higher levels for continuous improvement of information do call for robust use of high-quality evidence throughout the levels of the assessment systems.

One important development in educational assessment detailed in the ATC21S report is the increased ability, because of improved data-handling tools and technology connectivity, to combine formative and summative assessment interpretations for a more complete picture of student learning (Wilson et al, 2010b). The NETP describes how educators routinely try to gather information about their students' learning on the basis of what students do in class. We find that teachers in the classroom are already working with an enormous amount of assessment data that are often performance related. If good routes for transmitting information between classroom-based and large-scale settings can be identified, this could be a critical advance in the feasibility of measuring twenty-first-century skills in performance-based approaches.

It is not a luxury but almost a necessity to begin to combine evidence of practices in defensible ways if the goal is measuring twenty-first-century skills. Here, the availability of possibly very dense data may be key to effective practices, although such data density does not overcome the issues that are raised in this article concerning the need for evidence. However, the potentially available but currently relatively untapped evidence from classrooms, along with the vastly increased opportunities for efficient and effective data collection offered by the technology platform, mean much more evidence can be made available to understand student learning. This assumes, of course, that such data are collected in a way that maintains their status as evidence, and

that sufficient technology is available in schooling and perhaps even home and community contexts. Such practices would support the NETP's Recommendation 2.1, for designing, developing, and adopting assessments that give students, educators, and other stakeholders timely and actionable feedback about student learning to improve achievement and instructional practices, without compromising the quality of the evidence.

Schools can work together, too, to bring reality to these innovative assessment practices that combine data accrued within schools.

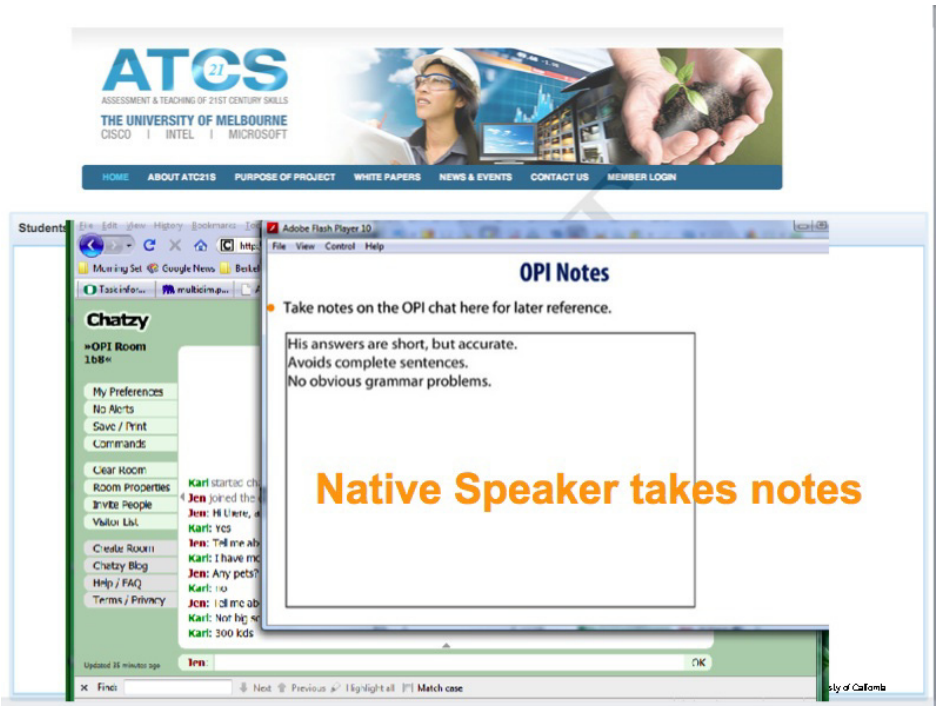
The Twenty-First-Century Outcome Spaces

Field testing is one part of the process that helps construct a defensible 'outcome space' or scoring interpretation for complex formats. Having a well-defined scoring interpretation that is supported by evidence should be part of the process of developing an item. Hence, it should be informed by research aimed at establishing the construct to be measured, and identifying and understanding the variety of responses students give to that task. In the domain of measuring achievement, a National Research Council (National Research Council, 2001) committee has concluded that for scoring, as well as for the previously mentioned issue of item development, a model of cognition and learning should serve as the cornerstone of the assessment design process. This model should be based on the best available understanding of how students represent knowledge and develop competence in the domain. It may be fine grained and very elaborate or more coarsely grained, depending on the purpose of the assessment, but it should always be based on empirical studies of learners in a domain. Ideally, the model will also provide a developmental perspective, showing typical ways in which learners progress toward competence.

The outcome space, to be fully useful, must also be exhaustive. There must be a category, or interpretation related to the construct, for every possible student response. As the NETP points out, some performances are so complex and varied that we do not have automated scoring options for them at present. In such cases, technology makes it possible for experts located far apart to provide students with authentic feedback. Constructed responses can be captured digitally as a by-product of computer test delivery and their scoring greatly facilitated whether scored by judges or by automated means. Online scoring networks (Mislevy et al, 2008) have been developed that can score constructed responses across time zones and by judges with different backgrounds in relatively short order. Automated scoring of written responses is a reality (Drasgow et al, 2006), and the automated scoring of speech is advancing rapidly (Zechner et al, 2009). Similarly, the automated scoring of some professional assessments (Braun et al, 2006; Margolis & Clauser, 2006) has been operational for some time.

The expectation of having an audience outside the classroom can be highly motivating for students, as the NETP describes. With the Assessment and Teaching of 21st Century Skills project, demonstration tasks, for instance, include not only peer collaboration but peer evaluation in the scoring model, as shown in Figure 1. We are currently researching how these practices can enter into model-based assessment of twenty-first-century skills.

Though the task remains challenging and is part of a research agenda to effectively incorporate some of these practices into robust assessments, the NETP report states that the United States is now 'acutely aware of the need to make data-driven decisions at every level of our education system on the basis of what is best for each and every student – decisions that in aggregate will lead to better performance and greater efficiency across the entire system' (US Department of Education, 2010, p. 25).



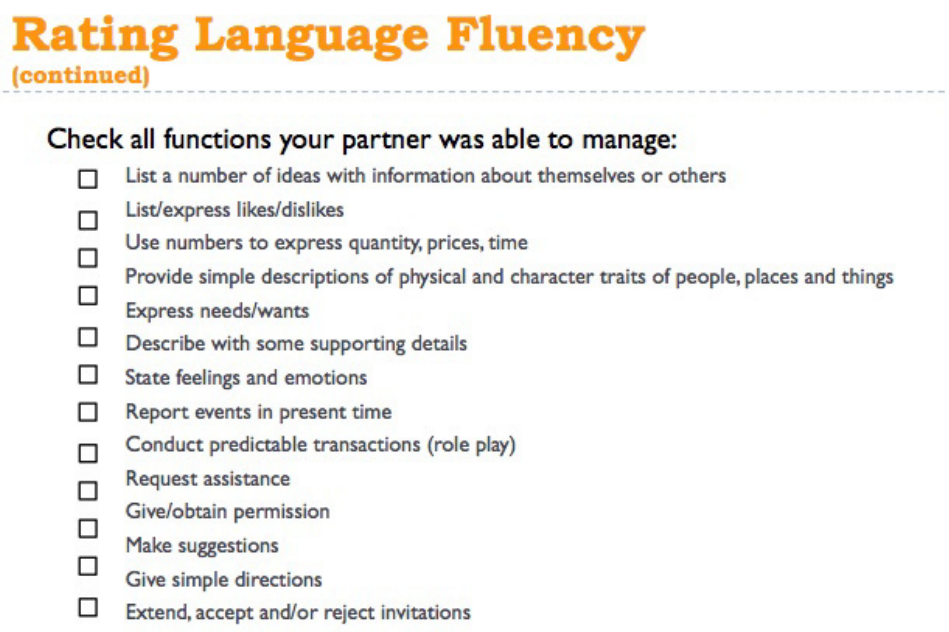


Figure 1. In this demonstration task, peer evaluation is included in the scoring model for communication. A native speaker and a language learner 'chat' according to a semi-structured interview protocol in a peer chat room, as shown in Figure 1a. The native speaker, who previously was trained in the rating system using scoring rubrics and pre-scored practice examples, then reflects on some communication aspects of the live conversation, as shown in Figure 1b. Compiling peer assessments from a number of interactions such as this with different partners is a 'wisdom of the crowd' approach to assessment and may provide some helpful measurement evidence.

Reporting of Results

Reporting of results gets back to the kinds of information that we aspire to provide. For the NETP goals, reports must be directly useful in enhancing instruction through targeted teaching of the skills being assessed. Ideally, we want these reports to provide timely and easily interpretable feedback to a wide variety of users, including students and teachers, parents and principals, administrative authorities and the general public. Finally, we want these reports to be valid and reliable by adhering to high technical standards in the development of the assessments and the analysis of the data.

Involving multiple stakeholders in the process of designing, conducting, and using assessment, as the NETP points out, is also important. Professional development for capacity building is critical. This supports the NETP recommendation that we need to build the capacity of educators and educational institutions to use technology to improve assessment materials and processes for both formative and summative uses. As good as technology-based assessment and data systems may become in time, given sufficient development of both the measurement technology needed and the information technology applications, educators will still need support in learning how to use them effectively in the instructional process. Use should not only be in terms of how to deliver the data collected from such assessments, but most importantly it should involve how to change teaching practices to effectively address what the data reveal, and how to engage in effective educational decision-making based on evidence of student learning.

The NETP points out that an important direction for development and implementation of technology-based assessment systems is the design of technology-based tools that can help educators manage the assessment process, analyze data, and take appropriate action. The NETP considers it a new role for assessments to diagnose how best to support an individual learner, and it may be true that this is a new perspective if taken as a role of large-scale assessments. The NETP discusses how one type of assessment – adaptive assessment – can facilitate differentiated learning (Scalise, 2007). It describes how this should not be confused with computerized adaptive testing, which has been used for years to give examinees different assessment items depending on their responses to previous items on the test in order to get more precise estimates of ability using fewer test items. The main difference described is in *use*, not necessarily in measurement functionality: the NETP describes how adaptive assessment has a different goal, designed to identify the next kind of learning experience that will most benefit the particular learner. In the process, however, getting precise estimates of proficiency quickly is possible and can be useful with this technique, especially since adaptive testing can be quite parsimonious and efficient in use of student time on assessments.

As the NETP states, many educators, parents, and students are concerned with the amount of class time devoted to taking tests (Perie et al, 2009). If teaching and learning are mediated through technology, and adaptive assessments are used to reach more precise estimates in less time, it may be possible to reduce the number of assessments and still provide plentiful information. With sufficient and appropriate measurement technology in place, data streams captured by online learning systems may be able to provide information needed to make judgments about students' competencies.

Universal Design for Learning (UDL) principles can also help make measures more accessible and precise through technology. The goal is to address the needs of students of greater diversity. This can include alternative representations for English language learner (ELL) students who may not be fully fluent in English, or offering other types of multiple modalities to accommodate more students and allow a broader portion of the population to show what they know and can do. Adaptivity may be useful in these practices.

The NETP describes the promise of how interactive technologies can support the measuring of complex performances that cannot be assessed with conventional testing formats. Such technologies, especially games, are described as also having the advantage of being highly engaging because they provide immediate performance feedback so that players always know how they are doing. The report calls for the Department of Education to provide a clearinghouse of information for states, districts, and schools about current research on and evaluation of new forms of technology-based learning and assessment.

Issues in Twenty-First-Century Assessment

Issue I: Context

While these formats are promising, key issues in cognition remain regarding using them to assess skills and knowledge such as collaboration, creativity, and communication. To take one example, the exploration of context specificity versus context generality is an exciting area of investigation, and it should not be seen as a barrier so much as an opportunity to explore and advance understanding. When defining constructs for measurement, a key question that can arise is the degree to which any particular context will influence measures of the construct. For instance, in assessments of vocabulary for reading, is the context of a passage selection important? For two respondents who may not fundamentally differ on the overall construct, will different instructional routes have led to different results when the testing is set in different contexts? What aspects of context may imply multidimensionality in the construct?

In traditional measurement, this has been a long-standing concern. Checks for multidimensionality and the examination of evidence for both content validity and internal structure validity are reflections of the issue. They may be addressed from a sampling perspective, by considering sufficiently representative sampling over the potential contexts to provide a reasonable overall measure of the construct. However, with twenty-first-century skills, the context can be quite distal from the construct. For instance, if communication is considered, it can be measured within numerous subject-matter areas. Communication skills in mathematics, involving a quantitative symbolic system, representations of data patterns, and so forth, is different from communication skills in, for instance, second-language acquisition, where mediation and meaning-making must occur across languages. On the other hand, some underlying aspects may be the same for communication across these contexts – it may be important to secure the listener's or audience's attention, monitor for understanding, and employ multiple avenues for understanding, across contexts.

At this point in the development of robust measures of twenty-first-century skills, there may be more questions than answers to the issue of contexts. Some educators see context as amounting to an insurmountable barrier to measurement. They may claim that the item-specific variability is so high that sound generalizations are not possible. Or they may go further and believe that there is no such thing as a general construct, just specific performances for specific contexts as measured by specific items. The means of addressing these concerns will necessarily vary from context to context, with some proving more amenable to generalization than others – that is, some contexts may be more unique than others.

Whether the context may alter proficiency estimates on the construct introduces questions of stability of the measures, and also calls for investigations of multidimensionality as these contexts are explored. Opportunities seem ripe to investigate commonalities of constructs across contexts, and divergences. Numerous methodological tools are available to consider the stability of constructs across contexts, and now may be an excellent time to do so, considering the nature of twenty-first-century needs for skills and knowledge across contexts.

Issue II: Continuous Improvement

The Continuous Improvement Data section of the NETP should be treated with caution. It describes how, once we have assessments in place that address the full range of expertise and competencies reflected in standards, student learning data can be collected and aggregated at many levels, for school accountability purposes as well as to improve learning outcomes and productivity. Since much of what is described in the report, such as having adequate constructs and showing evidence of measurement stability, remains a research agenda, it is ambitious to imagine having such approaches in place in the short term.

A key to understanding the proficiency status or knowledge status of students is the recognition of the intricacies of any testing data collected in the process. The Assessment and Teaching of 21st Century Skills methodological paper does advocate the reporting of results to users at all levels, from students and teachers up through school administrators and beyond. The ability to report such results, however, depends on a thorough comprehension of the sampling and

measurement design of the assessment, and its appropriateness for the construct, item design, scoring and reporting. Fair, valid and reliable measurement should be a foundational goal. We should not be distracted by whether online systems can deliver particular kinds of content and collect particular kinds of scores for a performance-based activity, but rather, we should be encouraged to use them as an opportunity to substantially improve measures, and thus learning outcomes.

Conclusion

In summary, creating the interconnected feedback system envisioned by the NETP could substantially help ensure that key decisions about learning are informed by data. Aggregating data and making them accessible at appropriate levels of the education system for continuous improvement is an important goal. The challenge associated with it is to combine the powers of technology with assessment practices that provide sufficiently robust evidence to support changes in practice effectively.

The complex terrain of twenty-first-century skills makes this no simple task. Information technology will need to employ robust use of measurement technology, which in some areas will require substantial further development. However, making the relevant data available to the right people, at the right time, and in the right form, as described by the NETP, is a natural direction to take for the information age in which we live.

Notes

[1] BEAR Center = Berkeley Evaluation and Assessment Center.

References

Almond, R.G., Steinberg, L.S. & Mislevy, R.J. (2002) A Four-Process Architecture for Assessment Delivery, with Connections to Assessment Design, *Journal of Technology, Learning, and Assessment*, 1(5). http://www.bc.edu/research/intasc/jtla/journal/v1n5.shtml

Binkley, M., Erstad, O., Herman, J., et al (2010) Assessment and Teaching of 21st Century Skills: defining 21st century skills. White Paper released at the Learning and Technology World Forum 2010, London.

Braun, H., Bejar, I.J. & Williamson, D.M. (2006) Rule-Based Methods for Automated Scoring: applications in a licensing context, in D.M. Williamson, R.J. Mislevy & I.J. Bejar (Eds) *Automated Scoring of Complex Tasks in Computer-Based Testing*. Mahwah, NJ: Lawrence Erlbaum.

Corcoran, T., Mosher, F.A. & Rogat, A. (2009) *Learning Progressions in Science: an evidence-based approach to reform*. New York: Center on Continuous Instructional Improvement, Teachers College–Columbia University.

Drasgow, F., Luecht, R. & Bennett, R.E. (2006) Technology and Testing, in R.L. Brennan (Ed.) *Educational Measurement*, 4th edn, pp. 471-515. Westport, CT: Praeger.

Kulik, J.A. (2003) *Effects of Using Instructional Technology in Elementary and Secondary Schools: what controlled evaluation studies say*. Arlington, VA: SRI International.

Margolis, M.J. & Clauser, B.E. (2006) A Regression-Based Procedure for Automated Scoring of a Complex Medical Performance Assessment, in D.M. Williamson, I.J. Bejar & R.J. Mislevy (Eds) *Automated Scoring of Complex Tasks in Computer-based Testing*. Mahwah, NJ: Lawrence Erlbaum.

Messick, S. (1989) Validity, in R.L. Linn (Ed.) *Educational Measurement*, 3rd edn, pp. 13-103. New York: American Council on Education and Macmillan.

Mislevy, R.J., Almond, R.G. & Lukas, J.F. (2003) *A Brief Introduction to Evidence-Centered Design*. Los Angeles: CRESST.

Mislevy, R.J., Bejar, I.I., Bennett, R.E., Haertel, G.D. & Winters, F.I. (2008) Technology Supports for Assessment Design, in B. McGaw, E. Baker & P. Peterson (Eds) *International Encyclopedia of Education*, 3rd edn. Oxford: Elsevier.

Moore, A.H. (2007) The New Economy, Technology, and Learning Outcomes Assessment: answering calls for change entails meaningful assessments of technology-enabled learning, *EDUCAUSE Quarterly*, 30(3), 6-8.

National Research Council (2001) *Knowing What Students Know: the science and design of educational assessment*. Washington, DC: National Academy Press.

Perie, M., Marion, S. & Gong, B. (2009) Moving towards a Comprehensive Assessment System: a framework for considering interim assessments. *Educational Measurement: Issues and Practice*, 28(3), 5-13. http://dx.doi.org/10.1111/j.1745-3992.2009.00149.x

Robyler, M.D. (2006) *Integrating Educational Technology into Teaching*. Upper Saddle River, NJ: Pearson.

Scalise, K. (2007) Differentiated e-Learning: five approaches through instructional technology, *International Journal of Learning Technology*, 3(2), 169-182. http://dx.doi.org/10.1504/IJLT.2007.014843

Scalise, K. (2010) The Influence and Impact of Technology on Educational Measurement. Invited Symposium at National Council on Measurement in Education (NCME), March 27, in Denver, CO.

Scalise, K. & Gifford, B.R. (2008) Innovative Item Types: intermediate constraint questions and tasks for computer-based testing. Paper presented at the National Council on Measurement in Education (NCME), session on 'Building Adaptive and Other Computer-Based Tests', May 1, in New York.

Scalise, K., Madhyastha, T., Minstrell, J. & Wilson, M. (2010) Improving Assessment Evidence in e-Learning Products: some solutions for reliability, *International Journal of Learning Technology* [Special Issue: Assessment in e-Learning], 5(2), 4-44.

Schum, D.A. (1987) *Evidence and Inference for the Intelligence Analyst*. Lanham, MD: University Press of America.

US Department of Education (2010) Draft National Education Technology Plan. http://www.ed.gov/technology/netp-2010

Wilson, M. (2003) The Technologies of Assessment. Invited presentation at the AEL National Research Symposium, 'Toward a National Research Agenda for Improving the Intelligence of Assessment through Technology', April, in Chicago.

Wilson, M. (2005) *Constructing Measures: an item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.

Wilson, M. & Sloane, K. (2000) From Principles to Practice: an embedded assessment system, *Applied Measurement in Education*, 13(2), 181-208. http://dx.doi.org/10.1207/S15324818AME1302_4

Wilson, M., Bejar, I., Scalise, K., et al (2010a) Assessment and Teaching of 21st Century Skills: perspectives on methodological issues. White Paper presented at the Learning and Technology World Forum 2010, London.

Wilson, M., Bejar, I., Scalise, K., et al (2010b) 21st-Century Measurement for 21st-Century Skills. Paper presented at the American Educational Research Association Annual Meeting, April 30, in Denver, CO.

Zechner, K., Higgins, D., Xiaoming, X. & Williamson, D. (2009) Automatic Scoring of Non-Native Spontaneous Speech in Test of Spoken English, *Speech Communication*, 51, 883-895. http://dx.doi.org/10.1016/j.specom.2009.04.009

KATHLEEN SCALISE received her PhD in quantitative measurement at the University of California, Berkeley in 2004. She is an assistant professor at the University of Oregon, in the Department of Educational Methodology, Policy and Leadership. Her main research areas are dynamically delivered content in e-learning, computer adaptive testing, item response models with innovative item types, and applications to equity studies. She recently served as a core member of the methodological group for the Assessment and Teaching of 21st Century Skills project created by Cisco, Intel and Microsoft, and for the Oregon state task force writing legislation for virtual public schools; she was also co-director of the University of California Berkeley Evaluation and Assessment Research Center (BEAR). In addition, she served with the Curriculum Frameworks and Instructional Resources Division of the California Department of Education. Her primary areas of work are in science and mathematics education. *Correspondence*: Dr Kathleen Scalise, 5267 University of Oregon, Eugene, OR 97403, USA (kscalise@uoregon.edu).

MARK WILSON is a professor in the Graduate School of Education at the University of California, Berkeley, USA. His interests focus on measurement and applied statistics. He has recently published three books: *Constructing Measures: an item response modeling approach* (Erlbaum), *Explanatory Item Response Models: a generalized linear and nonlinear approach* (Springer-Verlag) and *Towards Coherence between Classroom Assessment and Accountability* (University of Chicago Press –

National Society for the Study of Education). He is founding editor of the new journal *Measurement: Interdisciplinary Research and Perspectives*. Correspondence: markw@berkeley.edu