Creating Curriculum-Embedded, Performance-Based Assessments for Measuring 21st Century Skills in K-5 Students

American Educational Research Association
Vancouver, B.C.

Emily R. Lai

April 2012

ALWAYS LEARNING PEARSON

Abstract

This paper will share the author's experiences working with a large and diverse school district to design curriculum-embedded, performance-based assessments (PBAs) that measure 21st century skills in K-5 students. These assessments are designed to integrate seamlessly with classroom instruction; measure skills such as critical thinking, creativity, and collaboration in a developmentally-appropriate way; and provide information useful for supporting classroom teaching and learning. I begin by describing the context for the partnership. Next, I outline general principles that support the construction of the PBAs and explain the task development process. I also explore challenges in the development and validation of such assessments. Finally, I close the paper by speculating about future trends in PBA.

Keywords: performance-based assessment, 21st century skills

Creating Curriculum-Embedded, Performance-Based Assessments for Measuring 21st Century Skills in K-5 Students

The backdrop for this work is a partnership between Pearson and Montgomery County Public Schools in Maryland. This district is the 16th largest school systems in the United States and despite a diverse student body—both in terms of racial/ethnic identity and socioeconomic status—is nationally recognized for its outstanding student success. Winner of a 2010 Baldrige National Quality Award, this district boasts among the highest rates of attendance, AP participation, and graduation among similarly-sized systems. The partnership began in 2010, with the goal of developing *Pearson Forward*: a K-5 digital curriculum featuring embedded assessment and professional development resources. *Pearson Forward* is integrated on two separate levels: first, instruction in traditional subject areas (Reading, Writing, Math, Science, and Social Studies) is integrated via cross-curricular emphasis on a specific set of 21st century skills: critical thinking (encompassing analysis, synthesis, and evaluation); creativity (encompassing fluency, flexibility, originality, and elaboration); collaboration; metacognition; motivation; and intellectual risk taking. Second, assessment and professional development resources are integrated into the curriculum, rather than existing as separate entities, promoting coherence between what is taught, what is assessed, and areas for teacher support and improvement. The over-arching goal of the integrated curriculum is to begin the process of preparing students for college early on by focusing instruction on important 21st century skills and tracking student performance on specific benchmarks demonstrated to predict success in post-secondary education, such as advanced reading in grades K-2 and advanced math in grade 5.

As part of this partnership, Pearson is currently developing a series of PBA tasks to accompany the curriculum. The PBA tasks are designed to assess student proficiency in the specific 21st century skills targeted by curricular frameworks in the context of core academic domains. Thus, a PBA task might assess critical thinking in math, creativity in writing, etc. The tasks are also intended to support and be embedded in the curriculum, in the sense that assessment is virtually seamless with instruction. Finally, the PBAs are intended to be used for low-stakes, formative purposes: identifying student strengths and weaknesses, communicating with students and parents, and modifying instruction. The next section provides a more detailed explanation of the key features of the PBA system.

The Performance Assessment System

PBA Design

At each grade level, the curriculum is divided into 9-week instructional segments. During each 9-week segment, we design approximately 8 tasks to be administered sometime between weeks 6 and 9, with roughly two tasks designed to be administered per week. Each task is intended to be implemented over 1-2 class periods. Tasks are grouped into related families based on the particular 21st century skills indicators being assessed. Thus, for each 9-week instructional segment, there will be 2-3 PBAs, consisting of 2-3 tasks apiece. A complete task includes the task description, which provides step-by-step instructions for administration, student response sheets, teacher observation tools or checklists, student self-rating scales (if applicable), rubrics, and exemplars, which are similar to sample student responses. Rubrics are generic (i.e., not task-specific) and are holistic, rather than analytic. We have designed a single rubric for each 21st century skill at each grade level that describes student performance in terms of performance

levels rather than score points (e.g., “demonstrating,” “developing,” “emerging,” and “not yet evident”).

General PBA Principles

The integrated curriculum focuses on teaching and assessing several different skills or traits frequently classified as “21st century skills:” critical thinking, creativity, collaboration, metacognition, motivation, and intellectual risk taking. A recent review of the educational and psychological research literature connected to several of these 21st century skills suggests a set of general design principles that best support their measurement: (1) incorporating multiple measures to permit triangulation of inferences, (2) designing tasks of sufficient challenge or cognitive complexity, (3) including open-ended or ill-structured tasks, (4) embedding tasks in authentic, real-world contexts, (5) making student thinking and reasoning visible, and (6) exploring innovative assessment approaches that incorporate technology and nontraditional psychometric models (Lai & Viering, 2012). The PBA tasks Pearson is developing as part of this partnership incorporate several of these recommendations.

For example, each of the tasks evinces multiple measures of student proficiency, collecting some combination of the following types of evidence: student work products (e.g., pieces of writing, completed graphic organizers), teacher observations (collected via structured teacher checklists and/or observation tools), and self/peer evaluations, in the form of completed self- or peer-rating tools. Thus, a task assessing collaboration might consider the quality of the completed group work product, teacher observations about the student’s ability to work respectfully and productively with others, and the student’s self-reported collaboration skills and contribution to the group.

In addition, given evidence that many of these skills are somewhat domain-specific, we distribute tasks across content areas to permit more robust inferences to be made. For example, to account for the fact that skills such as critical thinking, creativity, motivation, and intellectual risk-taking tend to vary across domains due to varying levels of content knowledge, interest, and self-efficacy, we assess each skill in multiple domains. This strategy allows us to form a profile of student performance across domains, which is potentially informative to the teacher—not only in terms of the targeted 21st century skills, but also in terms of content. For example, a student demonstrating the creative thinking skill of fluency in writing, but not math, might need reinforcement of specific math knowledge and skills.

To assess 21st century skills for which it is critically important to create tasks of sufficient cognitive complexity or challenge (i.e., motivation, metacognition, collaboration, intellectual risk-taking), we generally design tasks that we determine to be just outside of reach for the typical student. In addition, we design tasks to be modifiable by the teacher to accommodate varying levels of student proficiency by offering suggestions for providing a greater level of scaffolding for students needing additional support or possible extensions for students needing additional challenge.

Mimicking an approach used by researchers studying intellectual risk taking, PBA tasks targeting intellectual risk taking generally offer several different versions of the same task varying in complexity and allow students to choose which task they will complete (see, for example, Clifford, 1991). For example, to assess intellectual risk-taking in reading, we designed a task in which first-grade students are asked to sequence and complete story boards and write accompanying captions to retell a story. Students could decide whether they wanted to retell a story that was very familiar to them or one that was much less familiar to them (i.e., had only

been read 1-2 times), with the choice of unfamiliar story arguably more of an intellectual risk than the choice of a familiar story.

PBA tasks within *Pearson Forward* are also designed to be open-ended. “Open” tasks allow students to decide what relevant information to use or how to use the information to solve the problem. These types of tasks require more metacognition and decision-making. In contrast, “closed” tasks are characterized by more teacher control and structure. With closed tasks, teachers indicate both the information to be used and what the expected solution will look like, and tasks typically emphasize a single, correct solution. “Open” tasks may also imply a higher level of examinee autonomy. In other words, open tasks may be differentiable by student interest or ability or permit students to make choices (Turner, 1995).

Researchers studying collaboration emphasize the importance of ill-structured tasks that cannot be solved by a single, competent group member (Webb et al., 1998). Ill-structured problems are those with no clearly defined parameters, no clear solution strategies, and either more than one correct solution, or multiple ways of arriving at an acceptable solution. Such tasks require knowledge, information, skills, and strategies that no single individual is likely to possess. Thus, PBA tasks focused on measuring collaboration typically provide a problem that would be very difficult for any one student to solve alone in the allotted time, which forces students to work together. For example, to assess collaboration in math, teams of 2nd -grade students were required to design and create a mosaic using multi-colored tiles and then to devise and implement a method for representing the data on tile color by creating a graphical depiction of it (e.g., a bar graph showing the number of tiles of each color used to create the mosaic).

Finally, tasks attempt to make student thinking and reasoning visible. Typically, this is accomplished by embedding some sort of informal teacher-student interview into the assessment.

As students are working, teachers are instructed to circulate through the classroom to observe and pose questions. For example, during a task designed to assess metacognition in science, Kindergarten students were asked work in small groups to construct a ramp out of common classroom materials (e.g., books, rulers, cardboard) and to test simple hypotheses about the effects of different ramp design features (e.g., height, surface material) on the distance an object would roll down it. Students were encouraged to share their thinking with teammates as they worked together. We provided a set of interview questions for the teachers to pose to individual students as they worked:

- How is it going?
- What are you doing right now?
- Why did you decide to build the ramp this way?
- What is working well about your ramp?
- What would you change about your ramp?

Using this tool, teachers could observe the extent to which students were able to share their thinking and explain their ideas to others, both key indicators of metacognition at the Kindergarten level.

Task Construction

Because tasks are intended to be embedded within the curriculum, they must closely reflect the content and processes emphasized during daily instruction. In order to achieve this level of integration, the process of writing PBA tasks relies heavily on existing learning tasks taken directly from the curriculum. Content specialists identify a set of learning tasks they believe would collectively represent important aspects of the 21st century skills indicators for that grade level. These learning tasks are “starters” for instructional activities and typically comprise

no more than 4-5 sentences. We transform them into assessment tasks by inserting more structure around desired student work products or performances, aligning student work products or performances more explicitly or transparently to the targeted indicators, suggesting a specific sequence of activities, and providing tools for collecting student responses. At the beginning of each task is a list of prerequisite knowledge and skills that students should have acquired to enable valid inferences about what they know and can do. Each prerequisite is linked to specific learning tasks in the curriculum that provide support for that skill, with the goal of enabling teachers to quickly verify whether they have provided adequate opportunities for students to learn the knowledge and skills and are thus ready for the assessment. Furthermore, because these tasks are intended to support and inform classroom instruction, there is an expectation that teachers will provide varying levels of student scaffolding as needed to enable all students, regardless of ability, to access the task. Thus, rubrics designed for scoring the tasks incorporate a consideration of the level and type of scaffolding provided to each student (e.g., whether teachers provide minimal, moderate, or extensive support and whether the support was related to skills that are the primary focus of the task or skills that are more ancillary to the task, such as writing ability in a task assessing creativity in math).

Challenges

In building this PBA system, we have encountered numerous challenges. First, it has been necessary to define 21st century skills in an operational sense, which entails identifying research-based behavioral indicators of these skills one might expect to see in K-5 students. Although some of these skills (e.g., critical thinking) were somewhat familiar to content specialists, others (e.g., metacognition, intellectual risk taking) were not. One of the first

challenges we tackled was to conduct a thorough review of the research literature on these skills to answer several questions:

1. How do researchers define them?
2. How do researchers measure them?
3. How do they develop?
4. How can you teach them?
5. How can you assess them?

Information gathered from this review of the literature supported the creation of written research summaries for each skill. These research summaries were shared with the entire Pearson team and with our partners in Montgomery County to support and facilitate discussions on the best ways to teach and assess these 21st century skills, thus providing a coherent approach across the program. We integrated these "lessons learned" into our PBA tasks as general design principles. In addition, we wrote up the research summaries for teachers and included them in the curriculum as professional development resources.

Once we felt we adequately understood the 21st century skills, a related challenge was determining how to build PBA tasks that adequately capture both cognitive and non-cognitive or affective aspects of student performance. Each of the skills targeted in the curriculum entails both cognitive and noncognitive or affective components (Blatchford et al., 2003; Clifford, 1991; Cross & Paris, 1988; Eccles & Wigfield, 2002; Facione, 2000; Sternberg, 2010). Cognitive components of these constructs include knowledge and strategies, whereas noncognitive components include attitudes, traits, and dispositions. To further complicate matters, cognitive and noncognitive elements are commonly confounded in practice. Thus, a student may be well-versed in a diverse set of problem-solving strategies, but be unable to use them consistently

because he cannot effectively regulate his emotions when he experiences challenges. To provide a more complete picture of student skills, both components should be addressed, ideally using complementary assessment modes (Ku, 2009). To assess motivation in reading, for example, we created a task that paired a teacher observation tool designed to capture students' use of strategic behaviors with a student self-rating tool designed to capture more affective aspects of motivation, such as the student's interest, self-efficacy, and goal orientation.

Another challenge is ensuring assessments collect student responses in ways that are developmentally appropriate for young children, who may have limited reading and writing capacity. To address this challenge, we use innovative approaches to collecting student responses, such as having students work with manipulatives (e.g., having students sort items into a graphic organizer) or by having students "tell" their responses to the teacher via a short student-teacher interview. In order to make this approach logistically feasible for the teacher, we often write tasks as though teachers will administer them to a single small group of students at a time, while other students are engaged in some other activity. This approach takes advantage of a common method for organizing instruction in lower-elementary classrooms—i.e., the use of "centers." Thus, tasks are designed to be administered in an environment that closely mimics the typical learning environment that students are accustomed to and students are allowed to respond in ways that allow them to demonstrate what they know and can do with respect to 21st century skills.

A third challenge specifically concerns the assessment of collaboration. As is well-noted, there are a number of challenges in assessing collaboration. Such challenges are related to the ways in which learning has traditionally been assessed in small-group settings. For example, educators are typically interested in obtaining individual student scores, but group assessment

tends to obscure individual contributions (Webb, 1995). Moreover, performance on group tasks is not necessarily representative of individual student performance, even when students turn in separate work products (Fall, Webb, & Chudowsky, 1997; Saner et al., 1994; Webb, 1993).

As Webb (1995) explains, assessments that occur in group contexts can fulfill several different purposes. For example, teachers may wish to determine how much a student can learn from collaborating with others, whether a group of students can complete a product together, or whether individual students can communicate respectfully with teammates. Group processes that support one goal may not support another goal. For example, if the goal is to measure a student's ability to learn from collaboration, then group processes such as co-construction of ideas, identification of conflict, giving and receiving elaborated help, and equality of participation should all be encouraged. In contrast, if the goal of group assessment is to determine whether a group can successfully complete a task on time, then group processes that facilitate student learning, such as trying to ensure equal participation among all group members, may be counterproductive. In this case, it will be more efficient to use processes that maximize group productivity, even if they minimize learning opportunities. Such processes might include letting the most competent student in the group perform most of the work (Webb, 1995).

Because the indicators of collaboration that we were interested in assessing include both group productivity and the use of helping behaviors, we design tasks that attend to both aspects of student performance. Thus, in a given collaboration task, the teacher checklist will include both minimal quality criteria for group work products, but also factors related to students' use of helping behaviors (e.g., communicating respectfully, soliciting diverse opinions). In addition, as mentioned previously, collaboration tasks are typically designed to require collaboration and

interaction among students, posing problems that no individual student is likely to successfully solve alone in the allotted time.

A final challenge is the inability to control factors related to the reliability of tasks. Because these tasks are essentially classroom assessments, the assessment developer has little control over administration and scoring conditions. Indeed, there is an expectation that teachers will modify the tasks to better suit the learning needs of their students. In addition, teachers, rather than trained raters, are scoring their own students' performance. Thus, generalizability and comparability of scores across tasks and teachers will likely not be on par with associated rates for PBAs in which administration and scoring conditions can be more standardized. However, as Lane (2010) points out, PBAs used for classroom, formative purposes do not require the same levels of standardization as are required for large-scale PBAs used for high-stakes purposes. In addition, we are employing a few strategies to help address the issue of reliability or generalizability. First, in writing tasks, we attempt to point out aspects of task administration that can be varied without impinging too much on valid construct measurement (e.g., allowable task modifications, variations in stimulus materials). Similarly, we point out aspects of tasks that should **not** be varied, such as the time provided to students to respond to prompts (when assessing the creativity indicator of fluency, for example). Second, professional development resources will include a series of tools designed to support teachers' administration, scoring, and use of PBA tasks in ways that are commensurate with the purpose of collecting information useful for modifying instruction. Tools might include online training modules addressing the purpose of the PBA tasks, using the rubrics, and interpreting results. Such tools are intended to minimize unwanted variability in administration and scoring conditions across teachers.

Future Directions

The coming years will likely bring several new directions in the use of PBAs. First, it is likely that the general classification of items and tasks as either "selected response" or "performance-based" will gradually be replaced by a system that more accurately reflects a diversity of open-ended item and task types. In other words, instead of a dichotomy, there is a continuum of assessment types and item formats varying in the extent to which student responses are structured. This continuum ranges from selected-response items (e.g., multiple-choice, true-false), to highly-structured constructed response items (e.g., short answer, fill-in-the-blank) to relatively open-ended and unconstrained types, such as live performances and portfolios. Indeed, we at Pearson are already at work on such a typology, which will soon be available online from Pearson's Center for Performance Assessment.[1]

A second trend in PBA is that educators increasingly expect assessments to provide actionable information that can be directly used to improve teaching and learning. This trend suggests that the next generation of assessments, including PBAs, should be capable of providing results that can inform teachers' next instructional steps. Thus, a teacher with a struggling student might be referred to additional learning tasks with similar demands as the PBA task on which the student can receive more practice. When PBA systems sample tasks from multiple domains, teachers can be provided with guidance regarding next steps for students with variable profiles (e.g., demonstrating the skill in some, but not all, subjects). Similarly, when PBA tasks collect both cognitive and non-cognitive information from students, teachers might receive guidance as to how to interpret results from these disparate sources. For example, if a student demonstrates a pattern of adaptive learning behaviors consistent with taking intellectual risks, but the student reports high self-efficacy and does not perceive his or her actions as particularly

risky, then the teacher can be prompted to provide that student with more challenging tasks and support for going outside of his or her "comfort zone."

Finally, another trend for the future of PBAs is the increasing use of technology enhancements, ranging from online administration of traditional PBA prompts to new item types that only exist in a computer format (i.e., immersive online simulations or games). Though the PBA tasks in this project are all paper-based, it is easy to imagine many of them being administered online. Still others could be administered in the classroom in a traditional fashion, with students formulating their responses on paper, but with teachers using digital devices to complete observation tools or checklists. Thus, perhaps the only limitation to increased use of technology to deliver PBA tasks such as those provided in *Pearson Foward* is the level of infrastructure necessary to support online assessment.

[1] Please contact the paper's author for the Center's url.

References

Blatchford, P., Kutnick, P., Baines, E., & Galton, M. (2003). Toward a social pedagogy of classroom group work. *International Journal of Educational Research, 39,* 153-172.

Clifford, Margaret M. (1991). Risk taking: Theoretical, empirical, and educational considerations." *Educational Psychologist 26* (3- 4), 263-297.

Cross, D. R. & Paris, S. G. (1988). Developmental and instructional analyses of children's metacognition and reading comprehension. *Journal of Educational Psychology, 80*(2), 131–142.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53,* 109–132.

Facione, P. A. (2000). The disposition toward critical thinking: Its character, measurement, and relation to critical thinking skill. *Informal Logic, 20*(1), 61–84.

Ku, K. Y. (2009). Assessing students' critical thinking performance: Urging for measurements using multi-response format. *Thinking Skills and Creativity, 4*(2009), 70–76.

Lai, E. R. & Viering, M. (April, 2012). Assessing 21st century skills: Integrating research findings. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, B.C., Canada.

Sternberg, R. J. (2010). Teaching for creativity. In R. A. Beghetto & J. C. Kaufman (Eds.), *Nurturing Creativity in the Classroom* (pp. 394–414), New York, NY: Cambridge University Press.

Turner, J. C. (1995). The influence of classroom contexts on young children's motivation for literacy. *Reading Research Quarterly, 30* (3), 410–441.

Webb, N. M. (1995). Group collaboration in assessment: Multiple objectives, processes, and outcomes. *Educational Evaluation and Policy Analysis, 17*(2), 239–261.

Webb, N. M., Nemer, K. M., Chizhik, A. W., & Sugrue, B. (1998). Equity issues in collaborative group assessment: Group composition and performance. *American Educational Research Journal, 35* (4), 607–651.