Top-Down versus Bottom-Up Innovation: The Relative Merits of Two Approaches to Designing Performance Assessments

Paper presented at the
Annual Meeting of the American Educational Research Association
Philadelphia, PA

April 2014

Ruth Chung Wei, Stanford University, rchung@stanford.edu
Susan E. Schultz, Stanford University, ses@stanford.edu
Daisy Martin, Stanford University, daisym@stanford.edu

Stanford Center for Assessment, Learning and Equity
Stanford University

The Stanford Center for Assessment, Learning, & Equity (SCALE) is an educational research and development laboratory at Stanford University's Graduate School of Education.

http://scale.stanford.edu
Littlefield Management Center
365 Lasuen Street
Stanford, CA 94305

Top-Down versus Bottom-Up Innovation:
The Relative Merits of Two Approaches to Designing Performance Assessments

Purpose and Significance

The high-stakes policy environment in which assessments have been used over the last decade have led to a heavy reliance on standardized selected-response tests designed and implemented by large-scale testing companies. Such tests have been criticized as being too far away from teachers' curriculum and instruction, and student learning (see e.g., Shepard, 2000; Shepard, 2003; Wiggins, 1990). On the other hand, teachers have been designing their own classroom assessments for generations, with test content closely tied to the taught curriculum. However, teacher designed tests have been criticized as being idiosyncratic, unreliable, and lacking comparability across teachers to be useful as credible measures. The performance assessment initiatives of the 1990s in which teachers designed their own tasks (e.g., the Vermont portfolio, the Kentucky portfolio system) were critiqued for the lack of comparability among teacher-designed tasks, among other issues, and weakened the validity argument for these systems, contributing to the discontinuation of their use (Koretz, 1998; McDonnell, 2004; Stecher, 1998). In more recent efforts to validate locally designed assessments (e.g., Rhode Island's Graduate Diploma System, Wyoming's Body of Evidence system, Student Learning Objectives-SLOs), the quality, comparability, and technical quality of the assessments used to evaluate students' competencies (Rhode Island, Wyoming) or to evaluate teachers (SLOs) have been raised as potential challenges (CEP, 2011; NRC, 2003; Goe and Holdheide, 2011).

The "Local Measures"[1] project, a performance assessment initiative implemented by the New York City Department of Education, deliberately sought to engage teachers in the process of designing summative performance assessments aligned to the Common Core State Standards (Literacy) and New York State standards in science and history, with the goal of combining both assessment design paradigms - designing assessments that are close to teachers' classroom

[1] "Local Measures" is a term that refers to one of the measures used to evaluate teacher effectiveness, as defined by the U.S. Department of Education in its "Race to the Top" state grant program. Race to the Top was established in 2009 through the American Reinvestment and Recovery Act, which allocated 4.35 billion to education programs. States awarded the federal grant, including New York, were required, as one condition of receiving the award, to create a teacher evaluation program in which value added measures of student achievement, along with other "local measures" such as observational ratings, professional development, and other student assessments, would be used to evaluate teachers.

instruction and curriculum, but with a design framework that included the involvement of expert designers, along with an iterative process for task review, piloting, and revision. This paper reports on the observations of the authors, who served as content and assessment design experts and facilitated the engagement of English, science, and history teachers in a design process that resulted in the development of performance tasks that potentially could be used to evaluate teacher effectiveness across the district.

In this paper, the authors reflect on the relative merits of two different strategies for engaging in the task design process, in both cases, within the policy framework and the teacher evaluation purposes discussed above. In the two years prior to the Local Measures project (autumn 2010 - spring 2012), task authors engaged in a "top-down" approach to performance task design in a different pilot that we call the "NYC Performance Assessment Pilot". In that project, experts in the disciplines (ELA, science, history, and mathematics) were engaged in designing performance tasks that were piloted and evaluated as potential measures of teacher effectiveness. Using this expert-design approach, tasks and scoring rubrics were first created by expert designers, vetted with educators, and refined for piloting. In the more recent "Local Measures" approach (autumn 2012 - spring 2013), expert designers met with "design teams" of teachers in their respective disciplines and grade levels to provide some basic professional development about task and rubric design, and to co-design the performance tasks and rubrics that would be piloted by the Design team teachers, as well as other implementing teachers during the same academic year.

The findings of this paper have important implications for evaluating the relative strengths and disadvantages of expert-designed versus teacher-designed performance assessments as valid measures of student achievement and teacher effectiveness, as well as the potential for different approaches to innovation in the field of assessment.

Theoretical Framework

"Top-down" design approaches versus "bottom-up" design approaches have not typically been discussed in the field of assessment development. Instead, top-down and bottom-up frameworks have been used widely to discuss divergent approaches to policy formulation and implementation, or in designing research approaches. These frameworks have been used in

policy research to explain the failure of many top-down education policies, however well-meaning, to be enacted in practice at the school and classroom level (Fullan, 1994; Cohen & Hill, 2001), to argue for the effectiveness of reform strategies that start with practitioners inside schools rather than from top-down mandates (Elmore, 2004), or to examine why bottom-up educational reforms are often impeded by policymakers (Honig, 2004). Fullan (1994) also critiques "bottom-up" or decentralized approaches to reform, and suggests that a combination of both approaches is most effective. The framework helps us understand the disconnect between policy and local practice and the importance of recognizing the central role of practitioners to support buy-in of policies, and implementing strategies that build capacity, support practitioners, and empower local leaders to appropriately translate the policies into local action. Similarly, in social science research, the top-down, bottom-up framework is used to more thoughtfully design research that attends to local contexts and the perspectives of practitioners or research subjects in order to improve the validity of the study findings and to more adequately explain the failures or successes of particular policies (Sabatier, 1986; McLaughlin, 1993).

In this paper, we apply the top-down versus bottom-up framework in a new way -- to describe two different approaches to performance assessment design and to explore their relative strengths and weaknesses. We also explore the efficacy of these two design approaches within a context in which a new top-down policy (teacher evaluation) frames the purpose of the assessment design and piloting activities, which has a bearing on the results of the design approaches and on teacher acceptance of the policy.

"Top-Down" Assessment Design Approach. In the first approach to assessment design, experts in each of four disciplinary fields (English language arts, science, history, and mathematics) were employed over a two-year period (2010-11, 2011-12) to develop new performance tasks, or adapt performance tasks from existing tasks, to serve as potential measures of student growth in learning and achievement. In this project, which we call the "NYC Performance Assessment Pilot", the district's goal was to explore the viability of using scores produced by these performance tasks to evaluate the contribution that teachers make to student learning and achievement, i.e., to evaluate teacher effectiveness. The specific methodology by which those scored would be used (whether in a pre-test/post-test fashion, or an end-of-year to

end-of-year method) had not been determined. The goal of this project was to ascertain the feasibility of developing standards-aligned, developmentally appropriate, and comparable performance tasks, where scores from those tasks could be used as a measure of student growth and achievement over time.

Performance tasks ask students to apply a range of knowledge and skills to perform and produce, which allows such assessment formats to go beyond correct and incorrect answers, and to measure deeper content understandings and disciplinary skills embedded in the Common Core State Standards, the Next Generation Science Standards, and the College, Career, and Civic Life Framework (National Council for the Social Studies). This is a big departure from the kind of tests that teachers have previously administered or the classroom assessments that most teachers have designed and implemented within their own classrooms. Moreover, the CCSS had just recently been released and adopted by the state (the NGSS and C3 Framework had not yet been published), so few practitioners had a sufficiently deep understanding of the standards.

Designing standards-aligned performance tasks that both meet the requirements of technical quality for high-stakes assessments and are developmentally appropriate, accessible, and engaging for students requires a level of expertise in performance assessment design that few classroom practitioners have. For this reason, the district contracted with SCALE, among other organizations with performance assessment expertise, to design and pilot performance tasks that would be entered into a task bank that would be made available to teachers citywide. Although designers did have considerable contact with teachers and solicited input and feedback from teachers and district personnel to shape the design of the performance tasks, the performance tasks were developed primarily through an **expert-design approach**, with the performance assessment expert taking the lead on design decisions for the development of performance tasks and scoring rubrics. Here, the main concern was with ensuring high-quality assessment designs that were comparable, aligned to standards, grade-level appropriate, and could produce valid and reliable scores. During the two years in which this design strategy was employed, large numbers of district teachers were involved in a pilot of these expert-designed tasks.

Based on the "top-down" versus "bottom-up" frameworks described above, we might predict that performance tasks designed using the expert-design approach might lead to less stakeholder buy-in and less enthusiastic participation from local educators, or we might anticipate breakdowns in the implementation of the performance tasks (with the tasks not being used as intended) because most participants were disconnected from the design process.

"Bottom-Up' Assessment Design Approach. Under the second approach to assessment design, instead of building more expert-designed performance tasks, the district sought to enlist the participation of teacher practitioners in the design of both the scoring rubrics and the performance tasks. This **teacher-involved design approach** was used in 2012-13, when the project was called "Local Measures". In the field of English language arts, approximately eight teachers in each grade span (elementary, middle, high school) were recruited from across city schools to serve as members of the "Design Team." These ELA teachers were joined by an even larger contingent of "Implementing Teachers", about 20 additional ELA teachers, to pilot their own performance tasks and provide feedback on the scoring rubrics. The main purpose of the Implementing Teachers was to explore the feasibility of implementing the assessment system on a larger scale. In the fields of science and history, approximately ten teachers in Grades 6-12 were recruited from across the city to serve as members of each "Design Team". Some of these teachers were specifically selected as representatives of the local teachers' union. Others were experienced teachers with some prior experience with performance assessment. Adoption of this new approach was likely a strategic one that sought to build capacity among district personnel and teaching staff to continue the performance task design work without the need to rely perpetually on expensive assessment designers. It was also likely a pragmatic and political decision to build support from the local teachers' union, whose approval was needed to move forward with the new teacher evaluation system.

The top-down vs. bottom-up framework suggests that this strategy of involving teachers as designers of performance tasks, in a context in which these assessments may potentially be used for high stakes purposes to evaluate teachers, might: 1) increase the validity of the assessments, 2) support user buy-in, 3) expand educator capacity in the design and use of assessments, and 4) make the assessments more instructionally useful.

This paper explores whether this "bottom -up" design approach and the involvement of teachers to contribute to the citywide task bank through the design of their own performance tasks, in fact, supports the first three goals. (We do not have enough empirical evidence to evaluate the fourth outcome about instructional usefulness.) We weigh the relative merits of the expert-design versus teacher-involved design approaches along three lines: 1) technical quality of the designed assessments; 2) teacher acceptance of these assessments as measures of their effectiveness as teachers; and 3) capacity-building of district personnel and teachers to design performance assessments.

Methods and sources

This paper draws on the experiences and observations of the authors, performance assessment experts who were involved in facilitating professional development and design studios with some of the "Design Team" participants in the Local Measures performance assessment project sponsored by the New York City Department of Education, as well as our relevant prior experiences in a two-year project ("NYC Performance Assessment Pilot") within the same district either as managers of the "expert design" approach or as the expert designer. From the NYC Performance Assessment Pilot, we have focus group transcripts, and responses to evaluation surveys completed by participants. For the Local Measures project, we did not have the opportunity to interview teacher participants and any survey data that was collected by the district was not shared with us. In this case, we rely on what we learned as participant observers during professional development events that we facilitated and reports from district personnel to inform our analyses.

Findings

I. Technical quality of the designed assessments

As noted above, the policy context in which the performance assessments were being designed and piloted was a high-stakes, top-down state mandate, in which the district was charged with establishing a teacher evaluation system in which "local measures" would be used as one component of the system. In the first two years, the district pursued an exploration of the feasibility of using performance assessments and in the third year, began gearing up for a city-

wide field test of the system by year four. In order to be useful as reliable and valid measures of student growth and learning, and to support high-stakes decisions, such as the determination of a teacher's effectiveness, even within a multi-measure system, there are several fundamental criteria for such measures -- a) they should have strong content validity - that is, they should be aligned with important disciplinary content and standards; b) they should be comparable in difficulty so that they can reliably capture student achievement and teachers' contribution to that achievement over a year long period; c) they should be equally likely to be connected to the curriculum taught by teachers and learning opportunities of students; d) they should be implemented in a standardized fashion to ensure that the students' scores are reflective of their own ability and achievement without assistance; and e) they should be scored in a manner that does not bias the scores (e.g., by students' own teachers) and so that there is some confidence in the reliability of scoring.

Table 1. Technical quality: Expert-designed vs. teacher-designed performance tasks

	Expert-designed*	**Teacher-designed****
Content validity	**Strong alignment** to CCSS and local content standards	**Variable alignment** to CCSS and local content standards
Comparability	**Consistent** in complexity, rigor	**Variable** in complexity, rigor
Connection to taught curriculum	**Strong alignment to state curricula** (e.g., Regents labs); Variable alignment to taught curricula (depending on teacher)	Variable alignment to state curricula; **Strong alignment to taught curricula** (individually designed for own class)
Standardized administration	**Variable** administration (tasks were partially curriculum-embedded)	**Consistent** administration (tasks were on-demand assessments)
Reliability of scores	**Sufficient rater reliability, variable task reliability**	**Unknown**

* with "Lead Teacher" and district review and input

**with expert designer/district review and feedback

Technical quality of expert-designed performance tasks

a) **Content validity**: The performance assessments produced under the "expert-design" approach were designed under highly controlled conditions and underwent extensive review both internally and externally by district teachers and other personnel. A great deal of care

was taken to select content that was aligned to the CCSS as well as local (state) content standards, as well as connected to curriculum typically covered in each grade level or grade span. Expert designers solicited and incorporated the input of teachers in selecting topics for the performance tasks during professional development and design sessions with teachers. Once the tasks were drafted, expert designers also solicited and incorporated feedback from teachers on the drafts. In science, expert-designed tasks incorporated key elements of the New York State Regents Labs (at the high school level) and labs required at the middle school level to ensure the assessments would align to the state curriculum. In history-social studies, topics were selected, with the advisement of teacher participants, that were likely to have been covered in the grade level or grade span, and were connected to units in the city's scope and sequence that would reasonably be taught during the administration window. In English language arts, the texts and topics were not necessarily connected to themes or ideas covered by teachers, but the tasks were designed to include all of the content and textual evidence needed to complete the task. Task prompts were revised and polished with the input of teachers, district personnel, and internal reviewers; texts and materials were carefully selected, adapted, and excerpted so that they were appropriate in complexity and accessibility; and teacher materials were developed to provide detailed and clear implementation guidelines.

b) **Comparability:** There is unclear evidence that the different expert-designed performance tasks piloted in this phase were psychometrically comparable, although certainly the intent was to produce tasks that were in the same realm of challenge and complexity, and to measure the same constructs. Because the performance tasks were piloted across the city across many teachers' classrooms, the district was able to collect information about how students performed on them across multiple schools and contexts.[2] In science, there were two types of assessments designed. The first type of assessment was designed to measure students' science investigation skills based on the NY State Regents Labs. The second type

[2] Scoring results indicated that students on average performed at the lower end of the score scale. We hypothesize that this was partially because of the nature of the school samples selected for piloting (the lowest performing schools), partially because teachers had just been introduced to the new standards and were unlikely to have begun implementing them, partially because students are unaccustomed to performance assessment, a form of assessment that requires more of them than simply selecting correct answers or regurgitating information, and sometimes because their teachers were not familiar with discipline-specific literacy and did not have the curriculum or instructional repertoire necessary to develop those skills.

of assessment measured scientific literacy skills requiring students to research the pros and cons of a science-related issue and to write an argumentative essay stating their position supported by evidence. Tasks within each type of assessment were designed to be comparable. Teachers had a choice of what type of task to implement in the fall (science investigation or science literacy), and in the spring, implemented the same type of task so that the same scoring criteria could be applied across both the fall and spring tasks. In history, all performance tasks within a given grade level were designed to be comparable in the types of disciplinary skills measured, e.g., ability to analyze sources, ability to understand multiple causation, and were scored using a common rubric with slight task-specific variations. In English language arts, all of the performance tasks required students to respond to a set of texts, whether excerpts of literature or informational texts, and to generate a final argumentative essay in response to a prompt.

c) **Connection to taught curriculum:** Teachers had limited choice about which tasks were to be administered. In most cases, two task choices were presented for each grade level, and teachers could select one of the two tasks to administer in both fall and spring of each year. As noted above, performance tasks were designed to be aligned to CCSS or state standards, as well as curricula that were likely to be taught in city schools. However, whether the performance tasks were aligned to curriculum that was actually taught by participating teachers is variable, depending on the curriculum used by individual teachers, and the extent to which teachers actually implement the state standards and the CCSS in their classrooms. It is likely that in classes in which there were Regents exams (at the high school level only), teachers were more likely to teach to the state standards and the Regents exams, as their students cannot graduate from high school without passing the Regents exams, and their own schools are held responsible for the test results through accountability policies. At the middle and lower grades, it is less likely that teachers are in "lock step" with the state standards, perhaps more so in the tested areas - mathematics and ELA - than in other subject areas (science, history).

d) **Standardized administration:** While detailed teacher materials were provided with performance task materials for students, the design of the performance tasks was conceptualized as being partially curriculum embedded -- meaning that teachers provide

instructional scaffolding to students as they completed the first part of the performance task, and then students completed the performance task under standardized, on-demand conditions. So, to some extent, teacher-level variability was designed to be part of the performance task administration. This makes sense for a measure of teacher effectiveness, because this allows the assessment to pick up variation in how well teachers are preparing and supporting students to complete the performance task. On the other hand, it may unnecessarily introduce another variable into the measure that cannot be controlled - effectiveness in task administration versus effectiveness of teacher instruction.

e) **Reliability of scores**: Scorer training of raters was facilitated by the assessment experts, and scoring was completed using a blind, distributed scoring method by an external contractor - a testing company. Generalizability studies (Cronbach, Gleser, Nanda, and Rajaratnam, 1972; Brennan, 2001) were used to evaluate the reliability of the scoring rubrics used to evaluate student work for the NYC Performance Assessment Pilot. Generalizability theory facilitates the investigation of how different sources of variation in the measurement design contribute to the reliability of scores. While the contribution of rater error to the total variance in scores is low (approaching .00 percent of score variance across most items), the overall reliability of scores across tasks and subject areas was variable. The science tasks had the highest levels of reliability, while English language arts tasks had the lowest levels of reliability. History tasks had variable levels of reliability, with most tasks falling below the levels of reliability that are generally considered acceptable (at least .80 with two raters). Analyses suggest that there were insufficient levels of reliability across performance tasks and rubrics to justify use of the tasks to measure teacher-level effectiveness using the tasks and rubrics designed for this pilot.

Technical quality of teacher-designed tasks

The performance assessments that were produced by Design Teachers in the third year were designed after they had completed a very modest amount of professional development on performance assessment (approximately two school days), using a set of tools, protocols, and guidelines provided by the facilitators. These included task templates and model/exemplar tasks with accompanying rubrics. While a small number of the Design Teachers had had prior experience as participants in the first two years of the NYC Performance Assessment Pilot, and

some of the Design Teachers had created and used their own performance tasks for classroom use, the majority of the Design Teachers had not had the experience of designing performance tasks to a particular set of specifications, or to be scored using a common rubric.

a) **Content validity:** After each cycle of task design (three cycles), teachers had the opportunity to receive feedback on their tasks from district personnel as well as the performance assessment experts. In science, the focus of the feedback was on clarifying the content to be assessed in the task, and including inquiry-based design elements that would require students to make choices and decisions instead of a step by step procedure. In English language arts and history, feedback included selection and adaptation of appropriate texts (content, quantity, length, complexity, and challenge), clarity in the wording of prompts, and supporting the design of prompts and selection of texts that do not bias students toward particular responses. Teachers used this feedback to refine their tasks before piloting. However, it was apparent that with each new cycle of performance task design, Design Teachers still needed substantial support to clarify the standards-aligned content to be measured, ensure that their task prompts were clear, sufficiently challenging, and appropriate in scope, and that the texts, stimuli, and materials they selected were appropriate for students as well as for the purpose of the task, allowing students to demonstrate the targeted skills and abilities.

b) **Comparability:** Performance tasks developed by Design Teachers were highly variable in content, the cognitive demands placed on students, the disciplinary skills needed to complete the tasks, and complexity of the texts and materials included in the tasks. Teachers designed these tasks with their own students in mind, and thus made design decisions, such as text selection, to support access to the task materials and the student product. In English language arts, a set of templates were used (the Literacy Design Collaborative writing templates for argumentative writing) to help teachers develop a more standardized set of prompts, but the resulting tasks were still highly varied in complexity and challenge. In science, the teachers relied heavily on the lists of sample prompts provided by the expert designer, but in many cases the teachers did not modify the prompts to ensure alignment to the goals of the task. In history, the expert designer provided a task template and then grade-level teachers modified that template in collaboration with the group and expert, but tasks varied widely in accessibility and coherence. The district employed "Common Core Fellows"

to evaluate the alignment of the tasks and text materials with grade level standards (CCSS), but these were personnel who did not have any expertise or training in assessment design. Design Teachers found their feedback to be more distraction than helpful, citing difficult to follow feedback, and a focus on elements that were not significant. We also found their feedback to be counterproductive at times and at odds with making the teacher designed tasks stronger. In science, the feedback of the Common Core Fellows was geared toward providing more scaffolding for students within the tasks, which often contradicted feedback from the expert designer.

The tasks that were produced by Design Teachers were vetted by the district initially through a review by the Common Core Fellows. Before granting approval, the district requested design experts to conduct a quality review of each teacher designed task, resulting in further revisions. Tasks approved by the district did become part of the citywide task bank from which teachers could potentially select tasks for implementation, so they had opportunities for pilot testing of those tasks in future years.

c) **Connected to taught curriculum:** The performance tasks developed by Design Teachers were often designed to be directly linked to the content and curriculum that they were using in their classes, and were piloted only by those teachers. In science, teachers focused on the Regents Labs and/or key concepts in science in an effort to represent what was taught in similar courses across the city. Teachers expressed concerned that the assessments should be fair to all their subject-area colleagues. (In a few cases, pairs or small teams of teachers worked together to produce a common performance task that they all piloted in common, but even in those cases, the teachers made adjustments to customized the tasks for their students.) However, given that the teacher designed tasks were designed for the particular classroom contexts in which teachers were working, and that they were piloted by one or two teachers, it is difficult to say how generalizable or applicable those tasks would be to other teachers working in other school contexts. In theory, because the tasks were connected to the curricula taught by Design Teachers, the tasks were more likely to be connected to curriculum taught in the district in general. However, we found that for some of the teachers, the curriculum was quite idiosyncratic.

d) **Standardized administration:** The tasks were designed to be 1-2 day (up to 90 minute) tasks, administered by teachers to their own students. The tasks are completed on-demand, with no assistance from teachers or peers. There was some debate among participants, particularly across grade spans, about whether the tasks should be administered like standardized high-stakes tests, or under low-pressure, low-stakes conditions. Teachers of early childhood and lower elementary students (K-5) were more in favor of low-pressure, low-stakes, un-timed administration conditions, not wanting their students to experience the performance task as a stressful event. On the other hand, teachers at the middle and secondary levels were more in favor of standardized testing conditions because they felt that it was fairer given the high-stakes purposes of the assessments (for teachers). They wanted their students to take the test seriously and to do their best. Implementation guidelines were designed and approved with teachers' input. These guidelines did call for more standardized administration conditions, with a focus on ensuring that the instructions for students were clear and accessible, that accommodations were made clear for students with disabilities and English learners, and that the results were fair.

e) **Reliability of scores:** During the pilot testing of the performance tasks in the Local Measures, teachers scored their own students' work following brief calibration sessions. The calibration sessions were intended less to build strong scoring practices and more to evaluate the strengths and weakness of the scoring rubrics for revision (which the teachers also contributed to in terms of design and content), to evaluate the fit between teacher designed tasks and the scoring rubric (with an eye toward revision of the tasks), and to consider revisions to the tasks and how they were implemented. In the English language arts teams, teachers scored their own students' work prior to even coming to these calibration sessions, and it is unclear whether those scores were even collected. In science, the district had teams of teachers who completed scorer training and met calibration standards scoring the student work using a blind distributed scoring method. The apparent plan (based on our last contact with teachers) was for these scoring teams to audit a small percentage of student papers to confirm the classroom teachers' score reliability. We do not have any access to the data to evaluate the reliability of local teacher scoring, but this approach raises questions about the fairness of using teacher generated scores for their own students' samples, especially in a high-stakes framework in which the scores are used to evaluate the teachers themselves.

Even with a credible audit system, there seems to be too strong an incentive to "game" the system in ways that would lead teachers to manipulate the scores.

II. Teacher acceptance of these assessments as measures of their effectiveness as teachers

During the two years of the NYC Performance Assessment Pilot (the expert-design approach), teachers were recruited to participate as pilot teachers to implement the performance tasks under varied conditions. In the first year, teachers from a group of low performing schools in the district (Schools Under Regents Review - SURR) were mandated to participate in the pilot. In the second year, "Lead Teachers" were recruited from across a number of different school networks. During these first two years, the expert designers had a substantial amount of contact with participating teachers. For example, the second year started with a four-day long summer institute, followed by other opportunities to meet with the expert designers, to provide feedback on the design of the tasks and rubrics, and to participate in scorer training and scoring events. However, in terms of task design, the role of the participating teachers was more of an advisory role, rather than as designers of the tasks themselves.

Participating teachers had varying levels of engagement in the project. They were invited to participate in professional development associated with learning about the performance tasks to support implementation and to build their capacity to teach in ways that would support their students' success on the assessments. Teachers and instructional leaders did share instructional strategies and resources focused on disciplinary literacy and Common Core aligned competencies, but the amount of time was short, attendance at these events were low, and this professional development time with the design experts was scaled back over the two years. Teachers were also given limited choice about which tasks that they were to administer in their classes, which took approximately one week to administer in some cases (less time for mathematics).

Despite the limited professional development time and limited choice in tasks, history teachers still reported enthusiasm and interest in using and learning about such performance tasks, tasks that looked like the Regents document-based questions, but as they noted, required more worthy skills. Teachers did, however, consistently voice concern over integrating new tasks with the

existing Regents' mandates and the use of these tasks in teacher evaluations. In focus groups and surveys, some teachers also expressed concern about the poor quality of the work being produced by students, many of whom enter middle and high school with very low reading and writing skills, an impediment to successful performance on the tasks. This, in combination with their concerns about the use of the performance assessment scores for teacher evaluation, was the overriding and primary concern of the teachers. As such, we can surmise that their enthusiasm about the performance assessment pilot was not high.

Table 2. Teacher acceptance: Expert-design vs. teacher-involved design approach

	Expert-design	Teacher-involved design
Process	**Generally satisfied** with consultative role	**Highly satisfied** with role as designers
Products (Performance Tasks & Rubrics)	**Generally satisfied** with tasks; Concern for low achieving students	**Highly satisfied** with tasks
Teacher Evaluation Policy	**Strong concerns** about proposed use of performance task scores	**Strong concerns** about proposed use of performance task scores

In the third year, during our engagement with Design Teachers for the Local Measures project, we had a similar amount of contact time with teachers, if not less time. Over a six month period, we had about 20-24 hours of contact with Design Teachers. In the case ELA, we met with teachers four times spread out over six months. The first session was a two-day session (6 hour days), while the remaining sessions were half days or 2-3 hour meetings after school. Science and History Design Teams met four times over a period of four months, with each meeting being a full day (6 hour days). During these meetings, participants engaged in discussions about their perceptions of the purpose of the project and whether they felt it was a fair way of assessing their effectiveness as teachers. Many of them offered such observations and critiques without being solicited. A small handful of teachers raised insightful questions about the validity of the tasks that they were designing as measures of teacher effectiveness. For example, a history teacher questioned the validity of designing history tasks that were structured to measure the Common Core Literacy Standards, rather than the history standards. He felt that the tasks would not truly be a reflection of his effectiveness as a history teacher, but might reflect the effectiveness of his ELA colleagues more than his own. In other words, he questioned whether specifying that the

tasks should measure standards shared across the content areas (CCSS) could truly distinguish between the contributions that different teachers make to a student's learning and achievement, particularly when students have multiple teachers at the secondary level. This teacher, as well as others, struggled with the dilemma of knowing that while they wanted to be involved in developing the shape and content of the assessments that would be used to evaluate them as teachers, they were not comfortable with the whole idea of being evaluated in this manner. However, the majority of the teachers involved with us as Design Teachers seemed to be satisfied with the process, with the performance tasks they built, and their involvement as Design Teachers. They appreciated the opportunity to contribute to the design of the scoring rubrics, learn how to develop their own tasks, while building and deepening their understanding of the Common Core State Standards. Science teachers appreciated learning about the design process, clarifying misconceptions on science concepts, learning more about inquiry-based teaching, and the benefits of using a common rubric, scoring student work with colleagues, and reflecting on their assessment and instructional strategies. They expressed comfort with being able to develop tasks that were strongly connected to their curricula and topics that they felt were worthwhile for teaching and assessing. The process also gave them a sense of being in a leadership role in making decisions about future assessments developed for students and teachers across the city. Some of the science teachers expressed that their experience in the project was the best professional development they had received. Overall, our impression of teachers' receptiveness of the resulting products was that the assessments they designed were fair and reasonable ways to assess student learning and achievement.

While Design Teachers were generally enthusiastic about being involved in the process of improving the assessment system for students, they did not necessarily see how this would be a fair measure for evaluating teachers. Many teachers expressed concern about the fairness of an evaluation system that was based on their students' performance, given differences in student achievement levels and teaching conditions across school contexts. They believed the assessments would help them assess their students' content knowledge and skill level but they did not understand how that would translate to evaluating them as teachers.

In the long run, we do not know how much initial teacher involvement in the design of the performance tasks and scoring rubrics will have an impact on teacher acceptance of these assessments as measures of their effectiveness. When we completed the project, the only teachers who were piloting the tasks were those who had had some direct contact with us as Design Teachers (ELA, science, history) or as "Implementing" teachers who contributed to the pilot by providing feedback on scoring rubrics and products (ELA only). We do not know whether the thousands of teachers in the rest of the district are more likely to accept the Local Measures system because they know that teachers have contributed to its design. Our understanding of the current system is that teachers are required to select from a few options for task implementation from a citywide task bank, some of which were created through our first two years with the district and some by teachers during the Local Measures phase. We know that during the Local Measures phase, negotiations over the terms of the teacher evaluation system with the city's teachers' union stalled at several junctures, especially over the imposition of an observational ratings system that they had little participation in designing or shaping. The local teachers' union has a very strong influence in the district, and if they see the Local Measures performance assessments as fair and reasonable measures because they had a hand in shaping and designing it, there is a possibility that teachers in the district will accept at least that part of the system. However, given that even the Design Teachers involved in designing the performance tasks that are part of the citywide task bank were concerned about their use for teacher evaluation, it seems optimistic that teachers across the district, who had no involvement in the project, would be supportive of the purpose of Local Measures.

III. Capacity-building of district personnel and teachers to design performance assessments

During the first two years of the NYC Performance Assessment Pilot, few district personnel were involved in shaping the content and design of the assessments, with most serving primarily as project managers. "Lead teachers" were involved in shaping the content and design of the assessments in the second year. However, they were less involved in the actual design of the assessments. Therefore, we can say fairly confidently that the expert-design "top-down" approach to performance task design did not result in capacity building or distributing expertise of district personnel/teachers in the design of performance assessments. On the other hand, generally high ratings of the participants who attended design and professional development

sessions suggest that teachers did deepen their understanding of the Common Core State Standards, implementation of the Common Core through curriculum or instruction, as well as discipline specific instruction in science and history.

Table 3. Capacity building: Expert-design vs. teacher-involved design approach

	Expert-design	Teacher-involved design
District Personnel	Minimal learning	Modest learning
Lead Teachers / Design Teachers	Significant learning	Significant learning
Implementing Teachers*	Minimal learning	Significant learning

***** In years 1 and 2, NYC Performance Assessment Pilot, the Implementing teachers were those who were not involved in design work ("Lead Teachers") but participated in implementing performance tasks. In year 3, Local Measures, the Implementing teachers (ELA only) were those who designed and piloted their own tasks (with less expert review and feedback) and provided feedback on the scoring rubric created by the Design Teachers. Implementing teachers in year 3 had less interaction with expert designers, but had similar amounts of contact time with the project.

During the Local Measures phase of the work, the "bottom up" strategy of enlisting Design Teachers, hand-picked to represent teachers leaders from across the city, to shape the design and content of the performance tasks and rubrics, seems inherently a strategy designed to build capacity among district personnel and teaching staff in the city. The Design Teachers we worked with directly had minimal professional development time with us - two six-hour days at the beginning of our engagement, and then a few hours at a time on three additional occasions during the next few months. Our direct engagement with the Design Teachers lasted between four to six months for a total of about 20-24 hours of direct engagement. Despite the limited nature of our work with the Design Teachers, many of them expressed satisfaction with the process and the learning they had gained about performance task design, the Common Core State Standards, and strategies for supporting their implementation of the CCSS. The district conducted interviews with teacher participants, and while the actual transcripts were never released to us, the district conveyed that the Design Teachers were overwhelmingly positive in their evaluation of their experiences as participants.

The district was also intentional about building the capacity of its own personnel, and phasing out its reliance on the assessment experts. District personnel and Common Core Fellows attended most of the sessions with teachers. From the beginning, district staff members more actively engaged in the design phase of the work so that they might learn from assessment experts as well as shape the products. They also requested that we teach them about what comprises high quality tasks so that they could provide ongoing feedback to teachers without our direct services. While we did not always have confidence in the expertise of district personnel to execute such reviews, it is laudable that they sought to build their own capacity to do so. At the same time, there is often high turnover among district staff. During the three years that we were engaged in this work with the district, key project managers and personnel came and went every year of the project, including the last year. This level of turnover suggests that efforts to build capacity within the district office will require deeper investments in supporting the development of expertise among personnel.

Discussion

The dilemma of teacher involvement in the design process

Our organization, the Stanford Center for Assessment, Learning, and Equity, has long held a commitment to teacher involvement in co-constructing and designing performance-based assessment systems that are intended to be used by teachers with their students embedded in ongoing classroom instruction and assessment. We have been careful to customize assessment systems for the users of those assessments, with the purpose of supporting buy-in from users and stakeholders and also building the capacity of teachers to understand the principles of designing high quality performance assessments and effective use of those assessments. We have found from experience that dropping in assessments from the sky is a less effective strategy for supporting teacher engagement and buy in, especially because performance tasks take a longer time and more teacher skill to administer.

However, there are also inherent dilemmas associated with teacher involvement in a design process and a system that relies on teacher designed performance tasks. These dilemmas were evident in our work with the Design Team teachers.

Benefits of teacher involvement

On the one hand, the Design Teachers' participation in providing feedback on the common scoring rubric yielded a product with which most participants felt comfortable and reflected their input and consensus. After multiple rounds of drafting, re-drafting, testing, and analyzing the rubric over a six month period, the set of common rubrics across the K-12 continuum (6-12 in science and history) reflects a thoughtful, consistent design across grade spans, transparent language that translates the CCSS/NGSS/state standards into user-friendly terms, and the values of the design team teachers so that the rubric measures and weighs appropriately what they considered to be the most important criteria for evaluating the student products. The common rubrics also reflect the technical demands and purposes for which they were being developed - a set of instruments that can differentiate among different levels of performance and show student growth over time on a consistent set of measures. In addition, while the authorship of these rubrics were ultimately the responsibility of the assessment experts, the process of taking teachers through multiple rounds of use and feedback helped to deepen their understanding of the common rubric and expectations of the new standards, and to develop a common language around achievement in the specific discipline. In this sense, the Design Teachers' engagement in the design process was a learning process. It was also a learning process for the expert designers, who learned from practitioners in ways that improved both the content and usability of the assessments that were designed. Design Teachers' participation with us in multiple cycles of the design-pilot-revise cycles helped us to refine our understandings of features of design that work best in classroom assessments, as well as design protocols and templates that were effective in supporting quality performance task designs. This suggests that the "bottom-up" assessment design strategy truly does have an "up" benefit, in that the knowledge and local expertise of practitioners gets taken up by design experts to innovate assessment designs that are not only grounded in practice but also lead to better products.

Deepening content knowledge and content pedagogy. The Local Measures project allowed teachers a unique opportunity to learn more about their disciplines, e.g., to examine inquiry-based strategies for teaching science, analysis of primary sources in history, and to think about alternative ways to assess their students. Teachers viewed these sessions as opportunities to network with other teachers and to improve their practice. Teachers openly shared their

instructional dilemmas and content gaps to the expert designer who worked with the teachers to strengthen their content knowledge and to build in time within the assessment design workshops for teachers to share their expertise on instructional strategies with each other. This team of teachers developed a strong network during the assessment design sessions and many of these teachers still communicate with each other on a regular basis.

Working on designing a new common rubric empowered the teachers to have ownership in the process as well as required them to reflect on the gaps between their current teaching practices and the vision of instruction outlined in the new standards. Teachers shared their enthusiasm and anxiety about the new standards. This is best captured in one teacher's quote, "My teaching will need to be significantly different using the new framework… its exciting and overwhelming at the same time."

Consistently, the design teachers reported that scoring student work from performance tasks and analyzing it was the best professional development they had ever received. They had extensive conversations and questioned each other to verify the evidence they gathered from the student work when assigning a score using a SCALE protocol. These conversations often highlighted student gaps and misconceptions about specific content. Teachers were surprised about how much information they could learn from these types of assessments. A pair of science teachers commented during one session, "Our regular tests just tell me whether the student got it right or wrong, these assessments require students to explain their ideas and provides a better method for us to identify what we need to re-teach or review." The science teachers were initially reluctant to work on a common rubric to score their student work because most of them had used a task-specific point system in the past. After the scoring session, they realized that a common rubric would enable them to monitor students' progress over time because they would be designing assessments that would be trying to measure the same performance outcomes. "Okay, we get it… if we create new assessments aligned to the outcomes on the rubric, we will be able to track whether students improve over the year."

Challenges of teacher involvement

On the other hand, the learning curve is steep for some teachers. Learning something new takes time to learn, practice, reflect, and improve, and we had a limited amount of time with the Design Teachers. As expected when learning a new process, teachers had varying levels of success for a variety of reasons. There were some teachers in the Design Teams who improved their ability to select appropriate stimuli and write clear, engaging, rigorous, and standards-aligned prompts over the course of their engagement with us. One strategy that supported greater comparability in task prompt design (for ELA performance tasks) was the use of the Literacy Design Collaborative writing templates at the 9-12 level. These tools provide a basic and common structure for the design of ELA writing tasks aligned to Common Core expectations for Argumentative writing. The templates also force teachers to produce writing prompts that are succinct and clear in their expectations for writing. Providing high quality models to deconstruct and analyze is another important strategy for helping teachers recognize effective assessment design features. Last, providing clear guidelines and constraints on topic and text selection and other design decisions helps to support more comparable tasks.

In the Science Design Team, one effective strategy was exposing teachers to a variety of performance assessments and having them identify what content and skills were being measured in the assessment. This was a challenging activity for many of the teachers but at the conclusion of the assessment design sessions, these teachers were able to pinpoint with accuracy what was being measured (i.e., performance outcomes) and whether the desired outcomes, standards, and performance task were all aligned. However, being able to examine and critique the alignment of an assessment is only the first step in learning how to design rigorous and standards-aligned assessments.

While some of the teachers improved in their ability to design performance tasks, a few teachers persisted in making some unproductive design decisions even after multiple rounds of task development, feedback, and piloting. Even after pointed feedback from experts, those patterns continued to be evident. This was particularly true of more "veteran" teachers who believed that they had been designing and using standards-aligned performance tasks all along, and were reluctant to reconsider their design decisions. A peer review approach was used during some of our sessions to provide some feedback to task designers. However, because some teachers are

hesitant to offer critique to peers (unless there is a secure professional relationship among them), this approach is less effective in producing high quality tasks.

Weak content knowledge. In the science team, teachers who lacked strong content knowledge or had misconceptions about the content had the most difficult time creating quality performance tasks. Through interactions with peers and the design expert, some of these teachers were able to resolve their content gaps and went on to produce acceptable tasks to be piloted with their students. Other teachers ignored feedback from the assessment expert and could not see the content misalignment until they scored student work samples with a peer and the issues became clear during the analysis.

Lack of engaging contexts for students. Another challenge for some science teachers was generating engaging contexts for the tasks and writing prompts requiring students to make decisions instead of providing all the procedures in the assessment. Teachers self-reported that they use very traditional teaching strategies with students. During the assessment design sessions, the design expert built in time for the teachers to share teaching strategies so they could learn from each other. These teachers would benefit from learning more about inquiry-based teaching strategies and having opportunities to practice inquiry-based teaching with feedback from coaches.

We believe that in a high-stakes assessment system in which the assessments are at least partially designed by teachers, expert review will continue to be necessary to ensure that there is some consistency and comparability in the rigor and appropriateness of the task prompts and text selection.

Conclusion

Performance assessments designed for high stakes purposes. The higher the stakes associated with the use of any assessment, the greater the demand for standardization and comparability, even from educators themselves. If the performance tasks like the ones developed in this project

are to be used to evaluate student achievement and teacher effectiveness, it is unlikely that teachers would be able to design their own writing tasks and select their own stimuli like texts because of the potential for wide variability in task and text difficulty across teachers. In addition, the performance tasks would need to be administered under more standardized conditions.

This does not mean that there should be no role for teachers. We saw that there were clear benefits to teachers as well as to the quality and content of the assessments when teachers were more closely involved in the design process. It may be possible that teachers continue to participate in the process of developing performance tasks, and that the tasks developed by teachers are vetted by experts for rigor and appropriateness for the grade level (and even piloted to produce evidence of its difficulty level). The tasks would be refined based on the feedback from experts and from pilot results, and if approved for use, these tasks would be entered into an official task bank from which teachers may select the task to be administered to their students. In this system, teachers could still have a central role in the assessment system, while ensuring that there is a sense of fairness and comparability in how the system operates for all teachers. To improve potential connections to teachers' curricula as well as accessibility to students, the selection of tasks within the task bank should include tasks with a wide variety of topical and thematic foci, as well as a wide range of text types that may be accessible to different segments of the student population. Teachers might select tasks that are either connected in some way to their curriculum or based on their potential for eliciting interest and engagement among his/her particular student body.

Neither does this common task bank system preclude teachers from designing their own performance tasks for formative purposes. We would hope that teachers would be assigning the same kinds of performance tasks at other points in the year (besides the beginning and end) in order to support students' development and progress toward the new standards. Through repetitive use of the performance tasks, teachers can build their own instructional repertoire as well as students' ability to respond to richer, more authentic tasks that have been customized to the curriculum or students' interests in ways that are likely to prepare them well for the summative high-stakes tasks. In addition, the repetitive use of the common rubrics developed

through this project would helps teachers and students internalize a common language and set of criteria about what is expected in students' performance, and by extension prepares students for on-demand assessments that measure the same learning outcomes.

Acceptance as a measure of teacher effectiveness. The policy context in which this project unfolded cast a perpetual dark cloud over participating teachers' work. While it is laudable that the district leadership moved toward a more teacher-involved "bottom-up" approach, there were many missteps along the way. The long-term goals of the project needed to be communicated in a more coherent way and the work should have been rolled out in a more well-planned, long-term timeline so that teachers better understood their role in the process and the rationale for some of the decisions that were made in a top-down fashion. Teachers would need to be more carefully selected for the design work (perhaps through an application process), and they (along with district personnel) would need much more intensive professional development to become more skilled in designing performance tasks.

It is still unclear to us that teachers in a system as large as the NYC public schools, with a powerful teachers' union, would accept and buy into a system of Local Measures that evaluates teachers and involves the use of teacher-designed performance tasks. There is no research evidence that we know of that the performance tasks teachers developed are comparable and reliable measures, and that teacher scoring is a fair approach given the high-stakes consequences. But given these high-stakes purposes, involving teachers in a "bottom-up" design approach (BY educators FOR educators) seems the most appropriate way to build teacher acceptance. The "bottom-up" design approach also has the benefit of building capacity among educators to develop and implement rigorous, standards-aligned assignments with students. This teacher-involved approach to assessment design should continue to include expert designers to review and support teachers' work, as well as include an iterative cycle of design-pilot-revise to build and expand teachers' expertise and to refine the quality of the performance tasks, design tools, and design processes. Through this kind of partnership between teacher designers and expert designers, it is more likely that the resulting assessments will be connected to taught curriculum (so that students have the opportunity to learn what is measured), and that the assessments will be user friendly, compatible with current classroom contexts, and instructionally useful.

****

References

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Center on Education Policy (CEP). (2011). Profile of state high school exit exam policies: Rhode Island. Washington, DC: S. McIntosh.

Cohen, D.K. & Hill, H.C. (1998). *Instructional policy and classroom performance: The mathematics reform in California*. CPRE Research Report Series RR-39. Center for Policy Research in Education, University of Pennsylvania Graduate School of Education

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Elmore, R. (2004). *School reform from the inside out: Policy, practice, and performance.* Cambridge, MA: Harvard Education Press.

Fullan, M.G. (1994, September). Coordinating top-down and bottom-up strategies for educational reform. *Systemic Reform: Perspectives on Personalizing Education* (September, 1994). Retrieved from: http://www.michaelfullan.ca/media/13396035630.pdf

Goe., L. & Holdheide, L. (2011). Measuring teachers' contributions to student learning growth for non-tested grades and subjects. Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from: http://files.eric.ed.gov/fulltext/ED520722.pdf

Honig, M.I. (2004). Where's the "up" in bottom-up reform? *Educational Policy,* 18(4), 527-561.

Koretz, D.M. (1998). Large-scale portfolio assessments in the US: Evidence pertaining to the quality of measurement. *Assessment in Education, 5*(3), 309-334.

McDonnell, L.M. (2004). *Politics, persuasion, and educational testing.* Cambridge, MA: Harvard University Press.

McLaughlin, M.W.(1993). What matters most in teachers' workplace context? In Little, J.A. &

McLaughlin, M.W. (Eds.) *Teachers' work: Individuals, Colleagues, and Contexts* (pp.79-103). New York: Teachers College Press.

National Research Council (NRC). (2003). *Assessment in support of instruction and learning: bridging the gap between large-scale and classroom assessment.* Washington DC: Committee on Assessment in Support of Instruction and Learning.

Sabatier, P.A. (1986). Top-down and bottom-up approaches to implementation research: A critical analysis and suggested synthesis. *Journal of Public Policy*, 6 (1), 21-48. Retrieved from: http://www.jstor.org/stable/3998354 .

Shepard, L.A. (2000). The Role of Assessment in a Learning Culture. *Educational Researcher, 29* (7), 4-14.

Shepard, L.A. (2003). Reconsidering large-scale assessment to heighten its relevance to learning. In J. M. Atkin & J. E. Coffey (Eds.), *Everyday assessment in the science classroom,* pp. 121-146. Arlington, VA: National Science Teachers Association Press.

Stecher, B. (1998). The local benefits and burdens of large-scale portfolio assessment. *Assessment in Education, 5*(3), 335-351.

Wiggins, Grant (1990). The case for authentic assessment. *Practical Assessment, Research & Evaluation, 2*(2). Retrieved March 27, 2014 from http://PAREonline.net/getvn.asp?v=2&n=2