EDUCATIONAL MEASUREMENT:
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Fall 2012, Vol. 31, No. 3, pp. 38–47

Judgment-Based Scoring by Teachers as Professional Development: Distinguishing Promises from Proof

Gail Lynn Goldberg, *Gail Goldberg Consulting*

The engagement of teachers as raters to score constructed response items on assessments of student learning is widely claimed to be a valuable vehicle for professional development. This paper examines the evidence behind those claims from several sources, including research and reports over the past two decades, information from a dozen state educational agencies regarding past and ongoing involvement of teachers in scoring-related activities as of 2001, and interviews with educators who served a decade or more ago for one state's innovative performance assessment program. That evidence reveals that the impact of scoring experience on teachers is more provisional and nuanced than has been suggested. The author identifies possible issues and implications associated with attempts to distill meaningful skills and knowledge from hand-scoring training and practice, along with other forms of teacher involvement in assessment development and implementation. The paper concludes with a series of research questions that—based on current and proposed practice for the coming decade—seem to the author to require the most immediate attention.

Keywords: judgment-based scoring, teachers as raters, professional development, scoring training

For as long as teachers have been engaged as raters to score constructed response items on assessments of student learning, whether formative or summative, low or high-stakes, claims have abounded as to the value of that enterprise to teachers as a form of professional development. With all but a few states now embarking on the design, development, and implementation of new assessment systems under the aegis of two consortia—SMARTER Balanced Assessment Consortium (SBAC) and the Partnership for the Assessment of Readiness for College and Careers (PARCC)—issues and implications of teacher involvement in a number of possible ways warrant closer examination of what has, until now, been accepted as common knowledge—that such involvement serves as a professional development opportunity for teachers.

This paper draws together findings from various sources in an effort to examine claims of positive impact: these include research and reports over the past two decades, information collected in 2001 from a dozen state educational agencies regarding past and current involvement of teachers in the scoring process, and interviews with educators who had served a decade or more ago as scorers for the Maryland School Performance Assessment Program (MSPAP), that state's pioneering performance assessment program. This inquiry into the impact on teachers of participating in scoring large-scale assessments and in related activities reveals that support for the popular perception is primarily anecdotal and often superficial; that consequences in terms of instructional practice and effect on student learning are rarely documented; and that practical demands have typically (and, many would argue, understandably) taken precedence over opportunities for teacher growth and learning.

From this multi-faceted overview emerge a number of research questions that beg investigation and various recommendations that might inform the engagement of teachers not only in scoring student work but in other aspects of the assessment development and implementation process as well. Current knowledge and experience appear to warrant a stance of cautious advocacy for teacher engagement, until the opportunities for professional development through activities like scoring are transformed into genuine occasions for enduring learning and positive classroom impact.

Background: What the Literature Says About Scoring of Large-Scale Assessments by Teachers

Over the past two decades, most state assessment systems have included at least some constructed response items, if only at selected grades—typically one or more essays by means of which competence in writing can be assessed and less frequently items in areas other than English language arts (Editorial Projects in Education Research Center, 2011). Although a few states have been making inroads into the use of automated scoring of constructed response, the norm is still judgment-based scoring by human raters. In most instances, scoring of open-ended items (essays, brief and extended constructed responses) on large-scale assessments has been and continues to be contracted out to professional raters; however, there were notable exceptions in the 1990s and early 2000s (all since discontinued for different reasons)—among them the Maryland School Performance Assessment Program

Gail Lynn Goldberg, Educational Consultant, Gail Goldberg Consulting, 2766 Westminster Road, Ellicott City, MD 21043; gailgoldbergconsulting@gmail.com.

(MSPAP), The Washington Assessment of Student Learning (WASL), Kentucky Instructional Results Information System (KIRIS), the California Learning Assessment System (CLAS), and the Vermont Portfolio Program (VPP). For at least some period of time during which each of these large-scale assessment reform initiatives were operative, teachers participated (albeit in varying numbers) in the operational scoring of student work. At present, teacher participation in operational scoring is a feature of a very small number of state assessment programs. Various initiatives have been implemented as alternatives to such participation, however, to bring understanding of scoring back home to the classroom.

The somewhat limited research to date on the impact of the involvement of teachers in judgment-based scoring (primarily of writing, portfolios, and performance tasks) has led to generally favorable views towards, and endorsement of, the practice as a vehicle for professional development (see, e.g., Darling-Hammond, Ancess, & Falk, 1995; Falk & Ort, 1997; Gambell & Hunter, 2004; Goldberg, 1994; Sheingold, Heller, & Paulukonis, 1994; Sheingold, Heller, & Storms, 1997). Those endorsements are based primarily, although not exclusively, on examination of formative or low stakes summative assessment enterprises—ones in which teachers were encouraged to examine student work in a collaborative environment rather than in isolation and were not under pressure to score "against the clock." From the perspective of the teacher-participants, the benefits of scoring experience most often cited are the clarification of standards, identification of desirable instructional practices based on examination of student work, increased assessment literacy that can inform classroom assessment practice, and deeper appreciation of the manifold ways that students might successfully demonstrate what they understand and can do. From the perspective of instructional leaders like one former assistant state superintendent in Maryland, the involvement of teachers in the labor-intensive scoring process served to "get them to buy into it" (Diegmuller, & Moody, 1995).

Investigation by Goldberg and Roswell (2000) into what teachers "took back home" from the experience scoring performance tasks for MSPAP, a statewide assessment initially devised in 1991 (and in operation through 2002) to drive school and instructional improvement, led, however, to a more nuanced view. In spite of the wide endorsement of their experience scoring MSPAP tasks, teachers often struggled when attempting to apply that experience to performance-based classroom instruction and classroom assessment practice. Goldberg and Roswell concluded that, by itself, "neither state-mandated assessment nor even the opportunity to participate in evaluation of students' work is likely to create the desired differences in teacher thinking and practice envisioned in school reform" (p. 289). Curiously, others who have subsequently cited their research (e.g., Cizek, 2001; Darling-Hammond & Rustique-Forrester, 2005; Gambell & Hunter, 2004; Youngs, 2001) generally ignore this more tempered viewpoint. And at this point in time, over a decade since Goldberg and Roswell conducted their study of the impact of scoring experience on teachers, there remains, to borrow a phrase that William Mehrens (1998) used to describe the consequences of assessment, "much more rhetoric than evidence" about teacher participation in scoring (or in other aspects of test design, development, and implementation) as professional development. Just as the content and format of tests can distort instruction (Murphy, 2007), so too can the involvement of teachers in the scoring process if it narrows their focus excessively or leaves them to draw and act upon sometimes faulty inferences about elements of quality and ways they may be embodied in student work.

With the general shift away from open-ended items and performance tasks in favor of assessments that could meet the demands of NCLB, interest in scoring by teachers diminished at home, even as attention to this feature of assessments in high-performing nations intensified (see, e.g., Darling-Hammond & McCloskey, 2008). This changed with the lead-up to the submission of proposals for new assessment systems designed to address the Common Core State Standards and in the subsequent award of Race to the Top funds to SBAC and PARCC; a number of the nation's educational leaders weighed in, and they included among their recommendations that local assessment of performance tasks and events in which teachers are responsible for scoring student work (subject to monitoring and auditing of scores), be a significant component of the new tests (Darling-Hammond, 2010, 2011; Hirsh, 2011; Lazer et al., 2010; McTighe & Wiggins, 2011). Both of the consortia proposals identified opportunities—even in some instances the necessity—for human raters, along with the benefits that could be realized if these were teachers from the participating states. The idea is attractive, for both its fiscal ramifications, which presume the use of teacher professional days (Darling-Hammond, 2010; McTighe & Wiggins, 2011; Topol, Olson, & Roeber, 2010) and even more so for the anticipated positive impact on teaching and learning. Pending the release of more detailed information on training and follow-up to teachers' participation in scoring, the degree to which that impact might accrue within the context of the new assessment systems is uncertain. In the meantime, however, the more recent experiences of various states that have involved teachers in some way in scoring large-scale assessments provide a lens through which to examine claims for scoring as professional development.

Involvement of Teachers in the Scoring of State Assessments Since NCLB

Although the number and types of open-ended items requiring judgment-based scoring by human raters has changed in the past decade, some states have devised and continue to support ways to involve teachers in the scoring process for their current assessment system. Beginning with a search on the Education Counts database of states that include a component requiring rater judgment (Editorial Projects in Education Research Center, 2011), followed by a review of information available on those states' websites on the then-current configuration of their assessments and a number of telephone inquiries conducted in February–March 2011, led to the identification of a sample of convenience of a dozen states that involve teachers in different ways in the scoring process. In some instances states were included in this sample because of their long history of, and commitment to, teacher involvement. In other instances, respondents themselves identified other states whose practices involving teacher participation in scoring had influenced their own.

Subsequent investigation into ongoing and anticipated efforts to engage teachers was conducted through unofficial telephone interviews with staff from those state educational agencies and several of the nation's largest assessment contractors. After an initial query about the nature and scope

of teacher involvement in scoring, subsequent questions to each respondent were varied, with the goal of fleshing out practices and purposes of each state. These exchanges—more conversation than interview—highlighted the strong degree to which conventional wisdom regarding the rewards of scoring as professional development has endured since the notion was popularized in the nineties. They identified innovative ways to encourage teacher involvement and the rationale for doing so, and also highlighted various obstacles to effectively using the scoring experience to build capacity to enhance teaching and learning.[1]

Localized scoring of student assessments has been realized at the broadest scale by New York. For that state's assessment system, nearly all open-ended items (regardless of content area) are scored by teachers either at the building level or within a geographic region (New York City, with one-third of the state's student population, uses the regional model). According to Steve Katz, Director of State Assessment, Office for Standards, Assessment and Reporting, there is plenty of anecdotal information from teachers, administrators, content specialists and others that scoring is a good staff development tool (Katz, 2011). Teachers, instructional leaders, and policy makers share the general sentiment that the experience helps "bring curriculum to light." The shared belief that scoring is good professional development underlies the policy in New York of allowing two of the four professional development days to be used for scoring of the state assessments, although one may cynically regard this practice as a means of accomplishing the task at hand without the expense of substitutes. Katz suggests that judgments from scorers on open-ended items are less meaningful if not local ("just a score, not an impression that can guide instruction"). However, some districts question whether there are diminishing returns over time; that doubt, along with sensitivity to the trade-offs related to cost, loss of instructional time, and the reduced opportunities for discussion given the tight timeline for results required by NCLB, led Katz to volunteer that he didn't think New York "if starting again would use this model." This model has also been vulnerable to very public critiques of scoring criteria and the decisions that ensue (see, e.g., Campanile & Edelman, 2010). There has been no formal research on the impact of scoring on instructional practice in New York, and Katz opined that it would be nice to survey teachers to learn what they have specifically taken back to their classrooms, put to use, and found effective.

Although a contractor's staff is responsible for scoring constructed responses in all other assessed content areas, Nevada teachers score that state's writing assessment. Initially "promoted as a huge professional development piece," payment for this summer work in a central location is also clearly a motivation to participate (Mudd, 2011). Those teachers who score are surveyed each year on the value of the experience, the uses to which they anticipate putting the scoring experience, and what they envision sharing with fellow teachers. However, given limited available resources (both human and fiscal), the focus of review of questionnaire data until now has been on logistics—how the scoring sessions can be managed more effectively and training improved—rather than on the instructional implications of the scoring experience. It was reported that all surveys are read, and are held onto for several years, but no formal analysis has been done nor any follow-up offered, although the department representative consulted indicated that, "If I could, I'd offer a course on how you follow up on different features in students' writing. What's the direct instruction that goes along with this?"

In recent years, scoring training for teachers and the involvement of teachers in operational scoring has been facilitated through technology. Oregon, for example, a state whose teachers have long played an extensive role in various aspects of test development and implementation (participation on content committees, bias and sensitivity review panels and in item development, as well as scoring), is transitioning to online scoring this year. Using a hybrid model, there will be some scoring conducted at central locations and some distributed scoring (the use of online services to transmit student work samples electronically to—and acquire scores from—raters at widely dispersed locations including but not limited to their homes) by experienced teacher-raters, using a platform developed by American Institutes for Research. All teacher participants are volunteers and will be compensated for this work. That state also intends to capitalize upon technology by supporting web-based training to build upon teachers' scoring experience so they can translate what they learned into action—to provide what Oregon's Director of Assessment and Accountability, Tony Alpert, identified as real professional development (Alpert, 2011). Incentives, beyond the pecuniary one, for teachers to participate in scoring include the opportunity to obtain continuing education credit through various institutes of higher education, meet requirements to maintain certification/licensure, and earn advancement on a district salary schedule (Hermens, 2011).

Oregon's approach to scoring the state's writing assessment illustrates the tension between practical demands and professional development. The desire to engage as many new participants over time (50% turnover being the goal expressed by ODEs language arts specialist) must be balanced against the fact that the greater the number of new trainees, the greater amount of training time is required. The main goal, not unexpectedly, is to get as accurate scores as possible, and therefore to hire those teachers who were effective readers in the past. Nevertheless, there is acknowledgement of the importance of the sharing that occurs during the training experience, and anecdotal accounts from teacher scorers support the idea that scoring will positively inform classroom instruction. No data on ways Oregon teachers have applied or intend to apply what they have learned from scoring has been formally gathered, however, nor has there been any investigation into the impact on students of teachers with scoring experience.

In North Carolina, the opportunity for teachers to participate in scoring the state writing assessment has been facilitated through a recently implemented system for distributed scoring (currently only at Grade 10), prompting the question of how that experience may foster professional development. For the past three years, that state has recruited North Carolina teachers to fill 20% of the scoring positions (with the remainder going to a contractor's professional scoring staff), bolstered by a study conducted by the Department of Public Instruction that demonstrated that teachers' performance in terms of accuracy and productivity was on par with professional raters. While the department regards teacher involvement as a form of professional development, however, it also is acknowledged to serve a public relations function; in involving teachers in the new distributed scoring model, "the number one goal was to create 'buy-in'" (Kroening, 2011).

In an endeavor to capitalize on aspects of scoring that can be valuable to a wider group of teachers, North Carolina recently created the *Writing Instructional System*, an online writing and scoring repository with instructional modules for teachers and mechanisms whereby teachers can score and provide comments on student work and engage in electronic "back and forth" with them for formative assessment purposes. LEAs and schools have the opportunity to use the system for entering student work, entering scores, providing feedback, and managing student writing portfolios. Although according to Jim Kroening, Lead Testing/Accountability Consultant at NCDPI, "teachers had a hard time moving from an assessment mindset to an instructional mindset," data as of mid-January 2011 revealed that over two and a half times the number of student responses had been uploaded as the same time the previous year (Kroening, 2011). To encourage buy-in, teachers using the *Writing Instructional System* may document this as evidence of meeting standards included in North Carolina's new teacher evaluation system and earn continuing education credits for completing online modules. Kroening uses terms more commonly associated with assessment—validity, reliability, high standards—to describe this instructional resource, even though he is clearly proud that there is "no 'testiness' to it." Instead, it appears to be a model for how beliefs about scoring as professional development can be transformed into professional development built on the foundation of scoring.

Washington State has had a considerable history of teacher involvement in scoring that began with teachers' participation in the scoring of writing in 2001, under the leadership of the former superintendent, Dr. Terry Bergeson, and Assistant Superintendent, Greg Hall. Over the next few years, scoring in additional content areas was added until teachers represented part of the reader pool that scored open-ended responses in Grades 3–8 for mathematics and reading, 4, 7, and 10 in writing, and 5, 8, and 10 in science. Although teachers at present continue to be involved in rangefinding (the selection of responses that are later used as models to train scorers), budget constraints in Washington State led first to the reduction in—and then in 2009 to the elimination of—teacher participation in operational scoring. However, since scoring by teachers had been endorsed by former Assistant Superintendent for Assessment and Student Information, Joe Willhoft, when he joined Washington's Office of the Superintendent of Public Instruction in 2004, proponents of teacher participation in scoring may anticipate that he will continue to endorse that practice as Director of SBAC, the position to which he was appointed in October 2010.

The impact on Washington State teachers of the scoring experience can perhaps be best illustrated by their involvement in the assessment of writing. Seeking to capitalize on what the scoring experience might have to offer to teachers, teams of 40 teachers per grade level were recruited (and selected to ensure coverage across the state), to train and score onsite alongside an outside contractor's staff. Eager to draw upon what teachers gained from the experience, the state's Writing Assessment Specialist collected from those participants' pre- and post-training and scoring survey data each year, along with self-reported anecdotal data on instructional implications (Elliott-Schuman, 2011). According to reports from 2005 and 2006, on a scale of 0 (low) to 4 (high), teacher-participants at each grade—4, 7, and 10—made the greatest gains in the degree of knowledge about the scoring training process and the quality of scoring training procedures. A considerable gain in confidence explaining the scoring process to others was also reported, perhaps justifying the impression that the experience turned teachers into ambassadors for the program. At grade 10, among teachers who had not previously served as scorers, gains in knowledge were the greatest (often going from 1 to 4) in all areas: the above two as well as knowledge of the use of the writing standards, rubric, anchor papers, and practice sets. When asked how the experience would impact their instructional practice, responses included familiarizing students with the rubric, encouraging voice and creativity (since unique approaches are not penalized), departing from the 5-paragraph essay, emphasizing the implications of purpose for writing and the importance of developing ideas with specific details and examples, and aligning feedback more with evaluative criteria (Elliott-Schuman, 2005, 2006). Teachers also were in agreement that the state's assessment targets were reasonable and reachable, and that scoring is fair and reliable.

The Washington Assessment of Student Learning (WASL) also served as the basis for one of the few instances of empirical research on the effectiveness of scoring experience as a professional development activity, a doctoral dissertation by Mary Alice Heuschel (2004), now Superintendent of the Renton School District in Washington State. Heuschel measured the effects of scoring training in writing and mathematics, using a train-the-trainer model in which teachers trained by a scoring contractor conducted scoring training in their home schools. Her study identified an increase in the number of students meeting the standard and demonstrated that teacher involvement in scoring had a positive impact on student performance on the state test. Among Heuschel's conclusions was that translating the skills and knowledge underlying scoring to effect teaching and learning requires "courageous leadership, administrative support that valued the expertise developed, and time for teachers to be trained and mentored." (p 89).

Kansas, which at one time supplemented local scoring by teachers of the state writing assessment with a centralized audit by teachers of a subset of responses (10%), has in recent years shifted to scoring by teachers done entirely at local level. Whether these are teachers of English language arts only or also of other content areas, and whether only currently employed teachers or retired teachers as well, are decisions made at district level. Of greater import in terms of potential impact of scoring on instructional practice is the redeployment of resources by that state's department of education towards the development and implementation of the online *Kansas Writing Instruction and Evaluation Toolkit* (KWIET). Schools will have the option of using KWIET or paper-and-pencil for the Kansas Writing Assessment (required biannually until replaced by the SBAC assessment) but also for local periodic and ongoing classroom assessment. KWIET shifts the emphasis from summative assessment of writing for accountability purposes to formative assessment of writing to inform teaching and learning. While there is an online function to look at scoring drift and may soon be another to determine deviation from agreement (key features of large-scale assessment scoring training and monitoring), the fundamental goal of the system is to use assessment to drive growth in writing performance rather than merely mimicking rater training and monitoring (as do some online, web-based programs that are

offered by several test publishing contractors) to give teachers the opportunity to think and act like scorers. Besides a prompt bank, KWIET allows users to dig deep into the criteria for evaluating each trait, examine exemplars for each criterion, and access a database for teaching the skills, processes, and understandings associated with each criterion/trait. The tool also facilitates conversation between and among raters and writers. Asked about the origin of KWIET, the Language Arts and Literacy Consultant for the Kansas State Department of Education recalls musing, "Wouldn't it be cool if we had this online tool we could use?" (Copeland, 2011). No systematic data on teacher use and impact of KWIET has been collected as of yet, nor has there been any examination of the instructional implications of teacher-moderated scoring using this online resource.

Even from programs that had at one time engaged teachers as scorers but no longer do, like Missouri and Nebraska, the mantra still echoes that scoring is great professional development. The benefits typically cited, based again only on anecdotal information, are that participation made the learning targets clearer for teachers and their students, clarified the meaning and implication of standards, and stimulated teacher thinking about ways to help students show what they know (Foy, 2011; VanDeZande, 2011); however, those reported benefits emerged from more conventional models of training and scoring (ones that mimicked the practices used to train and monitor contractors' staff of professional raters), ones in which teacher-scorers worked elbow-to-elbow and had the opportunity to engage in conversation, if only during breaks, that could inform what they took back from the experience. Given that the positive effects of scoring experience for teachers is typically associated with the opportunity for teachers to engage in discussion among themselves—as part of training and informally—one statewide writing specialist in whose state scoring has gone in 2011 to an outside contractor's staff noted that "there's not much difference between using contractors' readers and teachers if scoring is conducted remotely" (Foy, 2011). That difference, one can infer, is the value-added of collegial discussion, a key difference between most large-scale scoring enterprises and the practice of moderated scoring of classroom and formative assessment.

While the scoring of their statewide writing assessment is conducted by professional raters hired through a scoring contractor, in the past decade several states have elected to offer their teachers first live, and more recently online versions of the same training and qualifying activities to which those professional raters must submit before embarking on operational scoring. Working with Pearson, for example, the Virginia Department of Education offers teachers an online, web-based program called *Understand Scoring* (formerly known as *NCS Mentor*) that delivers annotated anchors, practice papers, and verification sets to teachers via *Perspective*, a cross-platform system. Feedback from teachers in Virginia has suggested that use of this program has helped promote their understanding of student learning targets; however, this is acknowledged to be a limited target—specifically, those features of writing that are central to passage of the state's minimum competency test—and is not intended to serve as professional development on the teaching of writing (Robertson, 2011). This distinction is critical: knowledge of, and experience with, any given scoring methodology in no way guarantees that it will lead to teachers' becoming better teachers of writing—merely that they are likely to be better able to guide their students to passing scores. For anyone to propose otherwise—and to treat training on scoring writing as if it were professional development on the teaching of writing—is a phenomenon against which Hillocks (2002) had cautioned nearly a decade ago, and still begs attention when involvement of teachers in scoring is entertained.

Arizona also collaborates with Pearson to provide online scoring training via *Perspective's Understand Scoring* feature; particularly since the state eliminated direct assessment of writing at grades 3, 4, and 8 (still retaining the assessment at grades 5, 6, 7, and HS), this training opportunity is of value to those who wish to continue writing assessment at the district or school level. Pearson provides the platform for online delivery and technical support, but Arizona teachers participate in the selection of papers and creation of annotations (Beach, 2011; Young, 2011). While it was reported by a representative from that state's department of education that teachers say that the program provides "great professional development," no data has ever been collected on the impact of this experience, either by the state or by the contractor, nor are there any plans to do so in the foreseeable future.

Both South Carolina and Florida utilized *NCS Mentor* in the past as a resource to illuminate the scoring process for teachers, but neither has transitioned to a newer online program for scoring training. Although scoring of South Carolina's writing assessment is done by a contractor's staff, their work is informed by rangefinding done by South Carolina teachers, and the description by a department education associate of the impact of that experience mirrors descriptions elsewhere of the impact of teacher involvement in scoring; through discussion, teachers learn from each other, get to see beyond their own classrooms, gain greater insight into the evaluative criteria and its nuances, and can serve as ambassadors for the scoring process among their colleagues. As in nearly all other instances in which teachers have been involved in one respect or another in the scoring process, there has been no systematic collection of feedback regarding perceived or actual impact (Howard, 2011). Rangefinding for Florida's writing assessment is also conducted in-state by experienced educators. While there is an assumption that exposure to scoring would help instruction, there has never been any attempt at formal data collection to examine this assumption (Lee, 2011). The obverse effect—the positive impact on scoring of participation by teachers—is intriguingly alluded to, but not elaborated upon, in an external evaluation of the scoring of the 2010 FCAT Writing Test (Geisinger & Foley, 2010).

Comparison of these twelve states' varied experiences with, and approaches to, exposing teachers to scoring (see Table 1) makes clear one commonality: the dearth of data at present—particularly more formal rather than strictly anecdotal—about the impact of that experience and understanding on teaching and learning.

Looking Back: The Long-Term Impact of the MSPAP Scoring Experience

While consensus among teachers who have scored large-scale assessments of student learning is that the experience had a positive impact—typically in terms of making standards and performance targets clearer, providing useful models

Table 1. A Profile of Teacher Participation in Scoring Various State Assessments as of Spring 2011

State	Assessment type(s)[1]	Scoring activity(ies)[2]	Reach[3]	Evidence of impact on teaching and learning
Arizona	W	R, SOST (active)	limited	no data collected
Florida	W	R, SOST (inactive)	limited	no data collected
Kansas	W	OS (local)	broad	no data collected
		FAS (local)	broad	
Missouri	W	discontinued	moderate	anecdotal data only
Nebraska	W	discontinued	moderate	anecdotal data only
Nevada	W	OS (centralized)	moderate	post-scoring survey on
	CR	R	limited	impact; no formal analysis or follow-up
New York	W, CR	OS (local)	broad	anecdotal; no formal data collection or research on impact
North Carolina	W	OS (distributed)	moderate	no data collected
		FAS (local)	moderate	
Oregon	W, CR	OS (centralized and distributed)	moderate	anecdotal; no formal data collection or research on impact
South Carolina	W	R, SOST (inactive)	limited	no data collected
Virginia	W	SOST (active)	–	no data collected
Washington	W	R (OS ceased 2009)	limited	pre/post survey data
	R	R		gathered and reported on teacher participation in OS

[1] W = writing; CR = constructed response items in other subjects.
[2] OS = operational scoring; OST = operational scoring training only; SOST = simulated operational scoring training; R = rangefinding; FAS = formative assessment scoring. Note that SOST is further characterized as active (promoted) or inactive (no longer promoted); in all cases of SOST, no information on reach was provided.
[3] Range describes the involvement of teachers, from few (limited range) to many (broad range).

of student work across performance levels, and increasing teacher confidence in the instrument and resulting data—these benefits appear very closely linked to the particular assessment system and the standards underlying that system. Given the changes in assessment practice over the period of time that roughly corresponds to before and after NCLB, a legitimate question is what—if any—impact the scoring experience may have on teachers over time and in the context of major changes to what is assessed and how assessment of learning takes place.

In Maryland, which in 2002 replaced its decade-long performance assessment of Maryland Learning Outcomes and Indicators with the Maryland School Assessment (MSA) tests based on the Maryland Content Standards, institutional memory of teacher scoring and its impact has greatly diminished. There exists no database from which to identify, let alone locate, teachers who engaged in scoring MSPAP; however efforts by the author (the former MSPAP Scoring Lead) to contact teachers who had participated in scoring for one or more years led directly or by referral to the identification of a convenience sample of twelve former participants. Although perhaps purely a coincidence, all twelve had experienced significant professional advancement since their days as teachers involved in scoring MSPAP. Three are now school principals, one is an assistant principal, two are district-level instructional supervisors for English language arts, one is a consultant to an educational non-profit, two are in resource positions, two are instructional specialists, and one works in a leadership role for a content area professional organization. To each of the former teacher-scorers, two questions were posed: *Looking back at the experience after a decade or more, what impact on your thinking and practice as an educator can you recall scoring MSPAP had at that time? What—if any—impact has that experience had which has endured, in spite of changes to the state's assessment system in terms of both its content and format?*

All but one of the former teacher-scorers consulted had a generally to extremely positive regard for their experience scoring MSPAP, often identifying the same or similar benefits, which included:

- Increased understanding of the standards and objectives by highlighting the alignment between standards (and their indicators) and component activities in assessment tasks.
- Clarity in regard to the learning targets (what's expected) and a "raising of the bar," along with increased understanding of "what good looks like."
- Access to a window into what students at each instructional level were doing in their classrooms based on what they were able to do on the test.
- Enhanced confidence in, and better understanding of, the data.
- Greater facility in the process—and increased appreciation of the value—of looking closely at student work.
- Increased understanding of, and the ability to implement, content integration.
- Strengthened understanding of, and commitment to, performance-based instruction and project-based learning, and ability to utilize assessment tasks as models of classroom activities and classroom assessment.
- Reinforced the difference between writing brief constructed responses and writing as a complex and recursive process (real writing).
- Enhanced the ability to implement classroom assessment more naturally and seamlessly.
- Increased familiarity with key cue words and item/activity stems and structures.

- Supported professional interaction by allowing for deep level of conversation as professionals around particular topics.
- Built trust and respect among members of an instructional community.

Given that many of the immediate benefits cited are program-specific (i.e., based on using performance tasks to address outcomes and indicators no longer operative), one might expect that the impact of scoring on these former teachers would have greatly diminished in the years since the assessment was discontinued; however, that was not in fact the case. One respondent who is now the principal of an environmental-science–oriented elementary school attributed her deep and ongoing support for interdisciplinary curriculum and instruction to evidence from scoring MSPAP that student learn and perform better when content is integrated. Her counterpart in another district confided that, "not a year goes by where I don't talk about my experiences with MSPAP; they still apply." The consultant who now works with schools and districts nationwide attributed her knowledge of true job-embedded professional development to her experience scoring MSPAP. Nearly all of those consulted considered the scoring experience to be a valid and enduring form of professional development because, at least for them, the activity involved far more that reading a response and assigning a score. It involved "seeing through students' eyes," identifying and questioning patterns in responses, realizing that "there's more than one way to show what you know." These understandings have withstood the test of time and changes in assessment policy and practice.

Several of the former MSPAP scorers interviewed identified characteristics of "real" professional development activities that informed their experience. One recalled the opportunities for exchange among team members and with project leadership, and asserted, "If there's no dialogue, there's no professional development." The majority expressed a favorable recollection of the interaction fostered by the scoring training and operational scoring experience, during which teachers rose to the challenge of assigning scores accurately and efficiently while they were encouraged and enabled to go beyond the immediate task to "finesse their own craft." Another readily agreed that the scoring experience qualified because she could relate the activity to what she did and could change and improve instructional practice as a result. She felt that scoring enhanced her ability to identify areas of need—something that subsequent experience has demonstrated can be "packed up and taken anywhere."

The one former scorer who disagreed with the notion that scoring was good professional development regarded the experience as "pretty much production work, assembly line." The negative response from that individual may reflect the fact that he scored off-grade level (i.e., he was teaching at the elementary level at the time but was assigned to score Grade 8 assessment tasks). It is also worth noting that he scored in the last two years of the program, during which similar complaints about the limited opportunity for discussion and reflection surfaced in the media (Desmon, 2002; Schulte, 2002). He denied that discussion among team members ever took place about the impact of item wording or format on student response or alternative ways that students might be successful—those features of the scoring experience that all others interviewed described and praised. Although it is not possible to verify his complaints, they lend credence to the belief that poorly conceived—or perceived—assessment experiences do much to promote teacher cynicism (Swain et al., 2010), while opportunities for conversation, reflection and forging of connections to the classroom are hallmarks of scoring experiences which teachers regard as good professional development.

Teachers Learning About and Through Scoring: Questions and Recommendations

Teachers who have been exposed to scoring methodology and resources, whether through scoring training only or through a more extended opportunity to assign operational scores, clearly often find value in that experience. This much is true, regardless of the assessment being scored and the uses that it will serve. What teachers report learning from scoring has tended to center around the assessment itself, however, rather than on broader implications for instructional practice in the content areas and domains being assessed. Furthermore, with only a few exceptions, evidence on the impact on teachers of scoring experience is anecdotal and is based on teacher perceptions and feedback often gathered in less than systematic ways. What data there are consistently suggest that perceived gains in knowledge about a particular assessment (how tasks are structured, how evaluative criteria support consensus judgments, and what are valued learning targets, for example) far outweigh and are far more common than gains in pedagogical knowledge. Which gains can or should be the aim of involving teachers in scoring assessments, now and in the future, is a question that it seems prudent to address.

Until now, the assertion that scoring serves (or can serve) as professional development has tended to get passed along without considering the great variety of experiences through which teacher engagement in scoring is filtered: highly structured scoring sessions in which teachers are proxies for "professional raters"; scoring enterprises which have attempted to balance the conditions deemed necessary to obtain valid and reliable results in a timely and cost-effective way with sensitivity to the conditions generally regarded as necessary to ensure ownership and application of new skills and knowledge; and less structured scoring activities primarily intended to inform teaching and learning rather than acquire data for accountability purposes. With many challenging decisions ahead regarding scoring methodology and best uses of human and fiscal resources, all those involved in designing and implementing large-scale assessment need to pause to consider what roles teachers can and should have in determining the scores assigned to student work—indeed, what roles overall teachers can and should have in all aspects of test development and execution.

Questions of whether, in what ways, or why teachers ought to be involved in scoring are no less pressing when considered in the context of deep interest in, and commitment to, automated scoring of items/tasks in the assessments of student learning being developed by both Race to the Top assessment consortia. In a scenario in which use of scoring engines facilitates scoring the bulk of summative assessment components, human judgments will be needed to inform various possible scoring methodologies: regression, rule-based, or hybrid. As a myriad of questions and issues receive attention (as they must) during what is likely to be accelerated evolution of the "state of the art" of automated scoring, the possible roles

of raters and/or domain experts should not be ignored. If it turns out—as is likely—that automated scoring alone cannot be justified, then we may wish to heed the recommendation to "keep well-supervised human raters in the loop" (Bennett, 2011, 3). It will be important to evaluate any case that may then be made for engaging teachers as raters—rather than or in addition to professional readers.

Assuming that any use of automated scoring will ultimately reduce, but not eliminate, the need for human raters to score various types of constructed response items and performance tasks, many aspects of typical large-scale scoring projects are likely to need to be retained—some form of inter-rater reliability checks, monitoring of output, and examination of score data for evidence of rater drift, for example. If professional development for teacher-scorers is a goal, however, some changes in practice are clearly warranted. The scoring of student work by teachers need not simply replicate the process of scoring a high stakes assessment. Useful models are already available for balancing scoring with sufficient rigor to meet the demands of a high-stakes assessment and scoring that allows for deliberation and debate about instruction that is so critical to teachers' learning. For example, an approach taken in Canada referred to as teacher moderation (as distinguished from "moderated scoring," or verification of judgment-based scores) combines attention by teams of teachers to rigorous decision-making when evaluating student work with dialogue about instructional implications (Ontario Literacy and Numeracy Secretariat, 2007). Although teacher moderation is associated more typically with classroom and formative assessment, there are lessons to be learned from this approach to involving teachers in scoring that might be applied to new assessment systems in the United States. Many other promising practices can be found among the assessment systems of high-performing nations (see, e.g., Darling-Hammond & McCloskey, 2008).

Another model, this one from the United States, is a writing assessment system called the *Analytic Writing Continuum* (AWC) developed by the National Writing Project (NWP), which features a centralized scoring system that maintains technical rigor and quality while allowing for adaptations in support of teaching and learning to meet local needs and interests nationwide (Swain et al, 2010; Swain & LeMahieu, 2012). Although project leadership acknowledged that participating teachers spent considerable time "nose-to-the-grindstone" (Friedrich, 2011), the Scoring Impact Study conducted by the NWP using an online survey and interviews demonstrated that they became more knowledgeable about writing assessment—including classroom assessment practice—and about writing instruction as well. At present, AWC-based inquiry is moving beyond the goal of obtaining valid and reliable scores essential to the NWP research agenda to informing professional growth in the teaching of writing. Still, to date, only a relatively small number of teachers have formally trained and calibrated as scorers, and thus questions remain about if and how teacher learning through scoring can be brought to scale.

At this point in time, then, the answer to the question of whether or not participation in scoring is good professional development is not "yes" or "no," but rather, "it depends." It depends on whether conditions that teachers have repeatedly identified as critical are established and maintained, conditions such as a collaborative environment, the opportunity to raise questions and exchange ideas, access to ample models of student work and assessment items that elicited that work that can serve as classroom resources, and tools to help them unpack standards and scoring criteria in ways that make sense to themselves and their students. It depends upon whether the learning that can come from the experience of scoring is reinforced outside of the scoring site or away from the computer station where distributed scoring has taken place so that teachers have the opportunity to conduct a "reality check" on the evaluative judgments they continue to make and the inferences they go on to draw from the assessment tasks, rubrics, and samples of student work to which they are exposed during scoring. By labeling an activity "professional development," it may be possible to reallocate human and fiscal resources to get a job that needs doing done. But unless there is real and substantial "take away" for teachers that ultimately makes a positive difference to students in terms of what they learn and can do, it is questionable whether scoring should be labeled as such.

It is worth bearing in mind that the technology platform that will ultimately support both SBAC and PARCC will make possible the delivery of training modules not only on application of evaluative criteria for scoring purposes but the instructional implications of assigned scores, along with recommended strategies and resources that can help inform teaching and learning. Online training and scoring can and should be extended beyond and outside of operational scoring and subsidiary scoring activities such as rangefinding, to expose the instructional community at large to be benefits of systematically evaluating student work.

We must seek to supplement our beliefs and the predominantly anecdotal data on the impact on teachers of involvement in scoring by conducting research to augment our knowledge and address many still-unanswered questions, including but not limited to the following:

- What, precisely, do teachers learn from the experience of scoring student work, and under what conditions is that learning maximized and sustained?
- What relationships exist between teacher perception, practice, and measurable impact on student performance?
- Is there a threshold for, and/or limit to, scoring experience (either in terms of number of responses scored or the duration of that enterprise) as a vehicle for teacher growth and learning? How might data that address this question impact the design and implementation not just of scoring but of other forms of teacher involvement in assessment systems as well?
- How does teacher learning acquired through scoring, in its various facets and formats, translate into subsequent instructional practice to inform teaching and learning and what—if any—impact does that practice have on student performance?
- What—if any—differences in instructional practice and student performance can be discerned based on actual scoring experience as compared to training on scoring?
- How can documented benefits to teachers directly involved in scoring enterprises be delivered through the design and implementation of professional development opportunities to those who have not had similar experiences?
- How can a technology platform be utilized to best support teacher participation in a wide array of assessment development, and implementation activities, as well as—and in support of—teachers' professional growth and development?

Only by answering these and other related questions more fully can we determine what substance there really is behind the popular pronouncement that scoring serves as an effective form of professional development for teachers.

It may be well to heed the lessons of the 90s, when calls for assessment-driven reform acknowledged the need for professional development but underestimated what was needed (Shepard, 1995). The sustained support that would benefit teachers would seem to require that we engage in the following preparatory activities:

- Conduct ongoing, research-based development and refinement of resources that capture the skills, knowledge, and dispositions that those involved in scoring have acquired so they can be shared with others.
- Provide opportunities for teachers to collaborate in teams with peers within a professional learning community, engage in reflective inquiry, and assume a variety of leadership roles related to knowledge acquired through scoring.
- Ensure coherence by drawing connections, for both teachers and students, between classroom-based examination of student work and more formal scoring procedures.
- Engage the expertise of specialists in the design and delivery of professional development to ensure that all types of teacher involvement, including but not limited to scoring, reflect the best that we know about effective professional development for teachers.
- Make more systematic the gathering of teacher feedback both immediately following their scoring experience and at a later time, perhaps by building this into online delivery of student work–much like an electronic exit interview.[2]
- Draw upon exemplary efforts to use the scoring experience to inform teaching and learning (including, but not limited to, the work of the National Writing Project).

The new assessment initiatives on the horizon create the opportunity to shift the focus from professional development on scoring to professional development through scoring. This does not require a choice between the two, but rather, better understanding of the possible ways of parlaying what we know about one experience to inform the other.

Acknowledgments

I am grateful to all those from various state educational agencies who generously provided background information on programs that offered teachers the opportunity to learn about and engage in the scoring process and to the Maryland educators who shared their recollections of scoring MSPAP. Thanks as well to Sandra Murphy and Linda Friedrich who brought me up to date on research sponsored by the National Writing Project on the impact on teachers of scoring writing, and to Jay McTighe, Steve Ferrara, and Isaac Bejar, who each provided helpful feedback on various drafts of this paper.

Notes

[1]The accounts in this section of various states' approaches to teacher involvement in scoring reflect policies and practices in place in Spring 2011.

[2]This suggestion owes its inception to Steve Katz, Director of State Assessment, Office for Standards, Assessment and Reporting, New York State Education Department.

References

Alpert, T. (2011, February 11). Telephone interview; (2011, May 17, 23). Follow-up emails.

Beach, M. (2011, February 8). Telephone interview; (2011, May 9, 10). Follow-up emails.

Bennett, R. E. (2011). *Automated scoring of constructed-response literary and mathematics items*. Retrieved April 14, 2011, from www.acarseries.org/papers

Campanile, C., & Edelman, S. (2010, June 8). NY passes students who get wrong answers on tests. *New York Post*.

Cizek, G. (2001). More consequences of high-stakes testing. *Educational Measurement: Issues and Practice*, *20*, 19–27.

Copeland, M. (2011, February 24). Telephone interview; (2011, May 11). Follow-up email.

Darling-Hammond, L. (May 2010). *Performance counts: Assessment systems that support high-quality learning*. Retrieved on January 8, 2011, from http://www.ccsso.org/Resources/Publications

Darling-Hammond, L. (January 2011). *New-generation assessment of common core standards: Moving toward implementation*. Paper retrieved April 25, 2011, from www.acarseries.org

Darling-Hammond, L., Ancess, J., & Falk, B. (1995). *Authentic assessment in action*. New York, NY: Teachers College Press.

Darling-Hammond, L., & McCloskey, L. (2008). Assessment for learning around the world: What would it mean to be internationally competitive? *Phi Delta Kappan*. *90*, 263–272.

Darling-Hammond, L., & Rustique-Forrester, E. (2005). The consequences of student testing for teaching and teacher quality. *Yearbook of the National Society for the Study of Education*, *104*, 289–319.

Deigmueller, K., & Moody, M. (1995, August 2). Bucking a trend, Md. assessments emerge as model school-reform tool. *Education Week*, 12–13.

Desmon, S. (2002, February 1). MSPAP graders see inconsistent scoring. *Baltimore Sun*. Section 1, p. 1.

Editorial Projects in Education Research Center. (2011). Education counts: Accountability. Edweek.org. Retrieved April 8, 2011, from http://www.edcounts.org

Elliott-Schuman, N. (2005). *Grades 4, 7, & 10 Washington Assessment of Student Learning in Writing: 2005 Teacher Scoring*. Seattle, WA: Office of Public Instruction.

Elliott-Schuman, N. (2006). *Grades 4, 7, & 10 Washington Assessment of Student Learning in Writing: 2006 Teacher Scoring*. Seattle, WA: Office of Public Instruction.

Elliott-Schuman, N. (2011, March 7, 9). Telephone interviews; (2011, May 17). Follow-up email.

Falk, B., & Ort, S. (1997, April). *Sitting down to score: Teacher learning through assessment*. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Foy, E. (2011, March 11). Telephone interview; (2011, May 9). Follow-up email.

Friedrich, L. (2011, April 11). Telephone conversation.

Gambell, T., & Hunter, D. (2004). Teacher scoring of large-scale assessment: Professional development or debilitation? *Journal of Curriculum Studies*, *36*, 697–724.

Geisinger, K. F., & Foley, B. P. (2010). *Considerations on the validation of the scoring of the 2010 FCAT writing test*. Lincoln, NE: Buros Institute for Assessment Consultation and Outreach, Burros Center for Testing, University of Nebraska-Lincoln.

Goldberg, G. (1994). Learning the score: What teachers discover from scoring performance assessment tasks. *Teaching Thinking and Problem Solving*, *16*, 1, 3–6.

Goldberg, G., & Roswell, B. (2000). From perception to practice: The impact of teachers' scoring experience on performance-based instruction and classroom assessment. *Educational Assessment*, *6*, 257–290.

Hermens, K. (2011, March 3). Telephone interview; (2011, May 9). Follow-up email.

Hillocks, G. (2002). *The testing trap: How state writing assessments control learning*. New York, NY: Teachers College Press.

Hirsh, S. (January 2011). *Building professional development to support new student assessment systems*. Paper retrieved April 25, 2011, from www.acarseries.org

Heuschel, M. A. O. (2004). A study of training 7th grade teachers in scoring the Washington Assessment of Student Learning (WASL) in writing and mathematics (Doctoral dissertation). Seattle Pacific University, Seattle, WA).

Howard, J. (2011, March 3). Telephone interview.

Katz, S. (2011, March 11). Telephone interview.

Kroening, J. (2011, February 24). Telephone interview; (2011, May 9). Follow-up email.

Lazer, S., Mazzeo, J., Twing, J., Way, W., Camara, W., & Sweeney, K. (2010). *Thoughts on an assessment of common core standards*. Retrieved April 28, 2011, from www.ets.org/s/commonassessments/pdf/ThoughtsonAssessment.pdf

Lee, S. (2011, March 2). Telephone interview.

McTighe, J., & Wiggins, G. (January 2011). *Measuring what matters: Part II—An enhanced assessment system supporting meaningful learning*. Hope Foundation Newsletter. Indianapolis, IN: Hope Foundation.

Mehrens, W. (1998 April). *Consequences of assessment: What is the evidence?* Vice Presidential address for Division D, American Educational Research Association, San Diego, CA.

Mudd, B. (2011, February 9). Telephone interview; (2011, May 9, 10). Follow-up emails.

Murphy, S. (2007). Some consequences of writing assessment. In A. Havnes & L. McDowell (Eds.), *Balancing dilemmas in assessment and learning in contemporary education*. London, UK: Routledge.

Ontario Ministry of Education Literacy and Numeracy Secretariat. (2007). *Teacher moderation: Collaborative assessment of student work*. Toronto, Canada: Author.

Robertson, T. (2011, January 11). Telephone interview; (2011, May 10, 11, 13). Follow-up emails.

Schulte, B. (2002, February 4). MSPAP grading shocked teachers. *Washington Post*. B 01.

Sheingold, K., Heller, J. I., & Paulukonis, S. T. (1994). *Actively seeking evidence: Teacher change through assessment development* (Center for Performance Assessment Report No. MS # 94–04). Princeton, NJ: Educational Testing Service.

Sheingold, K., Heller, J., & Storms, B. (1997, April). *On the mutual influence of teachers' professional development and assessment quality in curricular reform*. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Shepard, L. (1995). Using assessment to improve learning. *Educational Leadership*, *52*, 38–43.

Swain, S. S., & LeMahieu, P. (2012). Assessment in a culture of inquiry: The story of the National Writing Project's Analytic Writing Continuum. In L. Perelman & N. Elliot (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward H. White*. New York, NY: Hampton Press.

Swain, S. S., LeMahieu, P., Sperling, M., Murphy, S., Fessehaie, S., & Smith, M. A. (2010, April). *Determining what "good" looks like.* Paper presented at the meeting of the American Educational Research Association, Denver, CO.

Topol, B., Olson, J., & Roeber, E. (2010). *The cost of new higher quality assessments: A comprehensive analysis of the potential costs for future state assessments*. Assessment Solutions Group. Stanford, CA: Stanford Center for Opportunity Policy in Education.

VanDeZande, J. (2011, February 14). Telephone interview; (2011, May 13). Follow-up email.

Young, R. (2011, February 9). Telephone interview.

Youngs, P. (2001). District and state influences on professional development and school capacity. *Educational Policy*, *15*, 278–301.