The Use of Teacher Judgment for Summative Assessment in the United States: Weighed in the Balance and (Often) Found Wanting

Susan M. Brookhart
Brookhart Enterprises LLC
Duquesne University

Paper presented at the Invited Research Seminar
Teachers' Judgments within Systems of Summative Assessment: Strategies for Enhancing Consistency
Oxford University Centre for Educational Assessment

June 20, 2011

The Use of Teacher Judgment for Summative Assessment in the United States: Weighed in the Balance and (Often) Found Wanting

Abstract. *Lack of trust of teacher judgment for summative assessment of student learning in the United States (US) has a long history, and has been documented in research studies. There are a few bright spots, most notably in the use of teacher judgment for assessment in writing. However, most of these uses of teacher judgment for summative assessment circumvent teacher professional judgment by using standard rater training methods.*

In the United States (US), teacher judgment for summative assessment has been weighed in the balance and often found wanting. As this thesis suggests, the US literature presents a mixed evaluation of the use of teacher judgment for summative assessment. In this paper, I first present a very brief description of U.S. basic (kindergarten through grade 12) public education. This sets the context for the studies in the second section of the paper. Then in the second section of the paper, I develop an argument for the thesis, based on a selective review of US literature.

The Context for Summative Assessment in the United States

The common school in the United States developed in the nineteenth century (Cuban, 1993). The schools were locally controlled and often had one teacher for all the students, who could range in age from six through fourteen years or so. The teacher was often a young woman. As Ingersoll and Merrill (2010) put it: "The teaching force was transformed into a very large mass occupation that was a relatively low-paying, temporary line of work, predominantly for young, inexperienced women, prior to their "real" career of child rearing (e.g., Tyack 1974; Lortie 1975)." Early summative assessments included grade reports to parents (SGB, 1840/1992) and end-of-year recitations (Haertel & Herman, 2005).

By the late 1800s, exposés about poor instructional practices, including a famous presentation at the 1897 National Education Association meeting by pediatrician Joseph Rice, caused much consternation. Muckraking reformers outraged readers with stories of harsh instructional practices and little evidence of learning. However, the prevailing view of good teaching at the time was establishing mental discipline, without particular regard for measurable learning as an outcome (Gamson, 2007).

As the technology of educational measurement began to develop, through the work of E. L. Thorndike and others, progressive educators began to buy in to external standardized testing as a way to demonstrate the ineffectiveness of old-fashioned schooling by measuring learning. As Gamson (2007, p. 23) characterized it, "Educational evidence, systematically collected, was the crowbar that pried back the lid on traditional 19th-century schooling" by showing up the failure of many students to learn. The movement toward collecting systematic, scientific evidence of the results of education became known as "the new science of education." The "new"

standardized tests were not high-stakes external examinations that substituted for teachers' grades, but rather were used as a check on them in research and evaluation studies.

Around the turn of the 20th century, local schools gave way to more regional districts. More and more, these districts included a high school in addition to the lower schools. No longer the domain of young, inexperienced women, schools began to be controlled by professional educators. In the 1930's, Ralph Tyler and his colleagues began planning the Eight-Year Study, to investigate the effects of using progressive education methods in high school. Part of this effort was the establishment of a set of principles for educational evaluation, based on defining appropriate objectives, establishing and delivering learning experiences, and then evaluating whether the objectives had been achieved (Haertel & Herman, 2005). As common as this objectives-based approach to educational assessment sounds today, it was revolutionary at the time, and has been a big influence in educational assessment in the US to the present day.

Tyler's framework acknowledged the summative functions of teachers' grading and reporting to parents. The "rigorous, comprehensive testing program" served broader purposes of program and institutional evaluation, individual guidance to students, and the public relations function of confirming claims for progressive education (Smith, Tyler, et al., 1942, cited in Haertel & Herman, 2005, p. 6).

From the 1950's to the 1970's, interest in curriculum-based assessment of learning objectives was taken even further. A widely used taxonomy of educational objectives (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956), behavioral objectives, mastery learning, and measurement-driven instruction (Bloom, Hastings, and Madaus, 1971) all pushed Tyler's principles of evaluating learning to the level of fine-grained, classroom-level lesson objectives. Teachers were the main designers and evaluators of these classroom- and lesson-level objectives.

In fact, up until the minimum competency movement in the 1970's, almost all summative assessment and grading in schools were based on teacher judgment. Thus despite the belief in the science of standardized testing, teachers were free to evaluate their students as they saw fit. External achievement tests were used for program and institutional evaluation, and teachers' grading was the summative evaluation of individual student achievement (Haertel & Herman, 2005).

In the 1970's, the minimum competency testing movement began as a reaction to a pervasive dissatisfaction with public education. Minimum competency testing in the "basics" of reading and mathematics were required for graduation in 29 states by 1980 (Haertel & Herman, 2005). These tests were external (not teacher-made or scored) to the classroom. The movement peaked as concern moved from basic skills to "higher-order thinking" and as it became clear that minimum competency testing lowered expectations to meeting minimum requirements.

In 1983, a national commission published "A Nation at Risk" (NCEE, 1983), a widely influential document that proclaimed US students were falling behind, especially in light of increasing need for science and technology education. The report advocated rigorous and measurable standards and high expectations, a commitment to both excellence and equity, and recommended state and local use of standardized achievement tests. The educational reform movement of the 1980s and

1990s became known as the "standards movement" (Porter, 1993). It was similar to the minimum competency testing movement in that testing (not teacher judgment) was the method for certifying attainment, but different from it in that the tests were to monitor high expectations. To identify those expectations, states developed state standards for that achievement and outcomes-based accountability policies. Standards, policies, and assessment programs varied from state to state.

Federal legislation (NCLB, 2002) made states accountable for reporting proficiency level on standards, as measured by standardized tests, to the federal government. This legislation, called "No Child Left Behind," required reporting percent proficient at the school level, so the stakes were higher for schools than for individual students. It also required testing all students (including 95% of students with disabilities) and reporting school results disaggregated by subgroups of students: economically disadvantaged students, students from major racial and ethnic groups, students with disabilities, and students with limited English proficiency. States meet these requirements with annual standards-based tests in grades 3 through 8 and once during high school. In addition, twenty-eight states require high school exit examinations, and end-of-course exams of high school subject area standards are becoming increasingly more common among states that do not require high school exit examinations (CEP, 2010).

State academic standards still differ from state to state. In June, 2010, the National Governors Association Center for Best Practices (NGA Center) and the Council of Chief State School Officers (CCSSO) released a set of state-led education standards, called the Common Core State Standards in English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects and Common Core State Standards for Mathematics. The text of these standards is available at http://corestandards.org/. The intention is that states will adopt these as at least 85% of their state standards. As of this writing, 43 states (out of 51) and 1 territory (out of 5) have done so. It is fair to say that US education is becoming more standards-based and more centralized in that regard. New assessment systems planned for the common core standards include end-of-year comprehensive examinations as well as through-the-year tests, and plan to include some computer-scored performance assessment (Forgione & Doorey, 2010).

This very brief description of some of the history of US public education shows two different kinds of summative assessment have operated simultaneously. (1) Teacher judgment in the form of grading classroom summative assessments and assigning report card grades based on them has been, historically, the summative assessment with greatest impact for individual students. This remains generally true today, despite the increase in importance of external exams. (2) External testing for scientific, program, and institutional evaluation purposes has been, since the early 1900s, the summative assessment with greatest impact for schools. This remains very true today, in fact, centralization of standards and external assessment for accountability seems to be increasing.

Teacher Judgment for Summative Assessment in the United States

The quality of teacher judgment has both been studied in the context of both of these kinds of summative assessment (classroom grading and external assessment for accountability). In both contexts, teacher judgment has been weighed in the balance and often found wanting. Where

teacher judgment is trusted for summative assessments, for example in large-scale writing assessment, conventional rater training methods are used. That is, teacher judgments are accepted because the teachers have trained to a criterion, not because of their professional expertise or as the result of professional conversations.

The argument unfolds as follows:

- "Doomed from the start" – Teacher judgment in summative assessment has historically been distrusted in the US, based on early research on grading practices.
- "Never got a break" – Research and practice regarding teacher judgment in summative assessment continued to focus on documenting the lack of quality of teacher judgment, instead of focusing on how it could be improved.
- "A few bright spots" – Teacher judgment in summative assessment in writing and in the Work Sampling System for young children has been found acceptable and is used today.
- "The one that got away" – One state, Nebraska, tried to be a dissenting voice in favor of teacher judgment for summative assessment accountability. The Nebraska STARS experiment began before NCLB and held on through that legislative change (2000-2008) until public sentiment in favor of a state test won out. STARS relied on teacher judgment for the selection, administration, and scoring of assessments for accountability.
- "Be careful what you wish for" – Ironically, the standards movement in the US has created an atmosphere of reform that may ultimately result in raising trust in teacher judgment in the classroom, for grading, since grading on standards of achievement is a logical corollary of large-scale standards-based accountability.
- "Weighed in the balance and found wanting" – A definitive conclusion is hard to reach, as so many things in US education change and so many stay the same. However, it seems reasonable to say that current glimmers of trust in teacher judgment have a long history of distrust to overcome.

The sections below take up this argument. For ease of reading, the phrase "teacher judgment in summative assessment" is shortened to "teacher judgment."

"Doomed from the Start"

Some of the earliest educational research in the US studied the reliability of teacher judgment of student work (grading examinations) and found it to be unreliable. In the early 20th century, several researchers, most famously Starch and Elliott (1912, 1913a,b) studied what they called the 'reliability' of grading. The percentage grading scale (0-100) was commonly used at this time. In the first of their studies, Starch and Elliott begin (p. 442):

> The reliability of the school's estimate of the accomplishment and progress of pupils is of large practical importance. For, after all, the marks or grades attached to a pupil's work are the tangible measure of the result of his attainments, and constitute the chief basis for the determination of essential administrative problems of the school, such as transfer, promotion, retardation, elimination, and admission to higher institutions, to say nothing of the problem of the influence of these marks or grades upon the moral attitude of the pupil toward the school, education, and even life. The recent studies of grades have

emphatically directed our attention to the wide variation and utter absence of standards in the assignment of values.

In their 1912 study, Starch and Elliott sent the same two freshman examination papers in English, from a high school in Wisconsin, to two hundred high schools in the North Central Association, and received 142 usable, graded sets of papers. They analyzed, separately, 51 from schools where the passing grade was 70 and 91 from schools where the passing grade was 75.

The exam in question was a constructed response exam with a combination of six open-ended questions about grammar, literature, and writing. By more modern standards of essay test question writing (Coffman, 1971), they would not be considered well-written questions, with characteristics known to lead to difficult-to-score variation in student responses. However, they were typical examination questions at the time, and they led to great variation in grading. An example of their presentation of results, one of many charts with which they dramatically illustrated the unreliability of teacher judgment, is shown below.

Fig. 5 is a composite chart showing the values given to paper A by the entire one hundred and forty-two teachers. The values

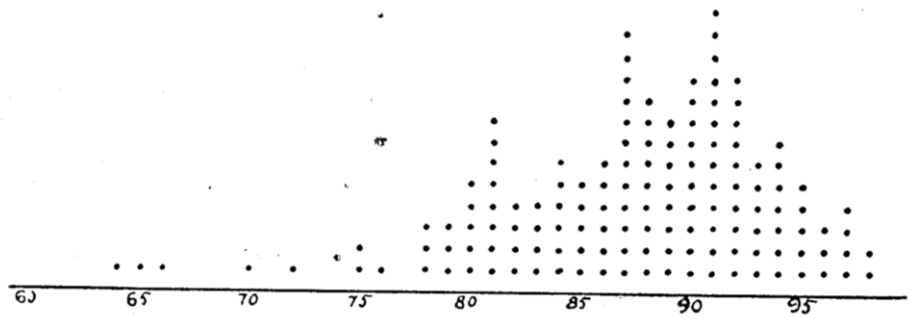

FIG. 5.—Paper A. Passing grade 75. Median 88.2. Marks assigned by schools whose passing grade is 70 are weighted by one point.

assigned by teachers in schools whose passing grade is 75 are represented as in Fig. 1, while the values assigned by teachers in schools whose passing grade is 70 are all raised one point, because the difference between the medians is approximately one point. Thus the combination of the two sets of schools in one chart is entirely fair.

Figure 1. Excerpt from Starch & Elliott, 1912, p. 451

Starch and Elliott replicated their study in mathematics (1913a) with a geometry exam and in history (1913b) with a United States history exam. Mathematics grading was even more variable than English. Variation in history grading was similar to mathematics. Starch and Elliott concluded (1913b, p. 680) that the probable error in grades for English, mathematics, and history was 5.4, 7.5, and 7.7 points (on the 100-point scale), respectively.

Starch and Elliott's analyses found four major factors contributed to the variability in grading (1913b, p. 681): "(1) Differences among the standards of different schools, (2) differences among the standards of different teachers, (3) differences in the relative values placed by different teachers upon various elements in the paper, including content and form, and (4) differences due to the pure inability to distinguish between closely allied degrees of merit." Arguably, only the fourth factor provides evidence to distrust teacher judgment. The first two are currently seen as more a matter for state standards and local curriculum reform, and the third a matter for advances in methods for writing test items or performance tasks with clear criteria. However, these 'grading unreliability' studies led to the move from percentage grading to letter grading, reducing the categories (and number of decision points) from 100 to 5.

These studies also set the stage for a distrust of teacher judgment of the quality of students' work, a theme which has characterized attitudes about teacher judgment in the US ever since. At about the same time, the "new science" of education swept in with the solution to the problem of unreliable teacher judgment: standardized, objective testing of student achievement.

"Never got a Break"

Teacher judgment in grading. Grading studies with a tone of "what's wrong with teachers' judgment," similar to the Starch and Elliott studies, never disappeared. They continued throughout the 20th century and continue into the 21st. Grading studies studies often compare teacher grading practices to the recommendations of measurement scholars, with mixed results that demonstrated teachers' grading of students' learning include many non-achievement factors. Some studies expressed the intent to show a need for research or teacher education in assessment practices (e.g., Stiggins, Frisbie & Griswold, 1989). Other times the expressed intent was more openly critical of teachers (e.g., Cross & Frary, 1999).

Studies of conventional teacher grading practices done with samples of US teachers consistently find that teachers add non-achievement factors into grades and produce unreliable, potentially un-interpretable grades. Besides achievement, teachers often consider students' ability, effort, and behavior in their grading decisions. Two reviews (Brookhart 1994; forthcoming) provide more complete reference lists of these grading practice studies. It is sufficient for purposes of this argument to show that studies of teacher grading practices continue into the 21st century and still find that teachers mix achievement and non-achievement factors when they make grading judgments. The "anti-teacher" rhetoric of some of the earlier studies has been softened, however, and researchers point out that achievement is the major, although not the only, factor teachers consider in grading (elementary, McMillan, Myron, & Workman, 2002; secondary, McMillan, 2001; both, Randall & Engelhard, 2010).

Teacher judgment and standardized tests. There is a small but definable literature on the accuracy of teacher judgment in general, when compared with "an objective measure of student learning," namely, an external standardized test (Hoge & Coladarci, 1989, p. 297). These studies conceive of correlations between teacher judgments of student achievement levels and test-based achievement as criterion-related evidence for validity.

Hoge and Coladarci's (1989) review of 17 studies (with a total of 55 correlations reported) published in the US literature between 1962 and 1988 concluded that teacher judgments of student academic achievement were generally accurate and valid, with judgment/criterion correlations ranging from 0.28 to 0.92, with a median of 0.66. Most of the 17 studies correlated judgment and test scores in Reading and Mathematics. Hoge and Coladarci (1989) also concluded there were individual differences in accuracy among teachers. Teachers were generally less able to judge the academic achievement levels of lower ability students than of higher ability students.

Harlen (2005) conducted a literature review for a similar purpose. She reviewed the international literature and found 30 studies published in English between 1986 and 2004, 11 with US samples. Only two studies were reviewed in both Hoge and Coladarci (1989) and Harlen (2005). In general, though, Harlen's conclusions agreed with Hoge and Coladarci's that teachers' judgments could be reliable and valid, although they were not in certain circumstances. She pointed out the value of detailed criteria in supporting accurate teacher judgments, and concluded with implications for assessment policy, practice, and research.

Ready and Wright (2011) investigated the reliability and validity of teacher judgment of students' early literacy with a national data set from the US National Center for Educational Statistics: the Early Childhood Longitudinal Study, Kindergarten Cohort (ECLS-K). Stratified random sampling targeted approximately 1,000 public and private schools around the US. The authors' final sample for this analysis contained 9,493 students in 1,822 classrooms in 701 schools, with sampling weights at both the student and school level. This sample can be called a representative sample of students, teachers, and schools in the US. Using hierarchical linear modeling, Ready and Wright (2011) found that about half of the variance in teachers' judgment of differences in literacy skills among children was explained by real differences in literacy skills. The remainder of the variance in judgment was due to various systematic errors. There was a small gender difference (favoring girls), and larger differences by race/ethnicity and socio-economic status. Inaccuracies in teacher judgment of students' literacy skills were more pronounced in lower-socioeconomic-status classrooms. These differences, especially the underestimation of low- socioeconomic-status children's early literacy skills, were large enough to be of practical significance. Finally, the authors reminded their readers that their study identified sources of inaccuracy in teacher judgment, but was not able to identify their causes.

Thus the literature studying the validity of teacher judgment of student achievement evaluated against the criterion of tested achievement "by no means constitute a ringing endorsement of teachers' assessment" (Harlen, 2005, p. 245).

Teacher judgment for assessment and accountability in "standards-based reform." The minimum competency testing movement in the 1970s relied mostly on basic skills tests. This "back to the basics" movement was critical of classrooms and, by association, teachers, and sought external measures of achievement instead of teacher judgment. Fremer and Dwyer (1977, p. 5) pointed out that those involved in minimum competency testing should think about ways that test based information could be combined with teacher judgments. Nevertheless, the minimum competency testing movement did not make a place for teacher judgment in the way the standards-based reform movement tried to do.

As the standards movement geared up in the 1980s and 1990s, so too did a call for more performance assessment. This was in part a response to the call for higher standards that included student thinking, and in part a response to the reform argument that performance assessments would be "tests worth teaching to." Vermont, Kentucky, California, and Maryland designed statewide assessment systems that included performance assessments. Teacher judgment in scoring these performance assessments was studied as part of the evaluation of these programs.

Vermont used portfolios for assessment in writing and mathematics in grades 4 and 8. Teacher ratings of portfolios were unreliable in both subjects in 1991-92; they improved somewhat in mathematics but not writing in 1992-1993. Teacher ratings did not differentiate well between best work and other work, nor between dimensions (the criteria upon which the work was rated). Ratings were unreliable enough that even aggregated to the school level, scores were not dependable (Koretz, Stecher, Klein, & McCaffrey, 1994). Vermont ultimately dropped their portfolio assessment system.

In 1997, the state of Ohio passed legislation prohibiting schools from promoting to fifth grade any student who did not meet a designated cut score on the fourth grade reading test, unless the student was excused from the test or both the student's principal and reading teacher agreed the student was academically prepared for fifth grade. Thus this exception allowed teacher judgment to override a test score. To investigate this policy, Cizek, Hirsch, Trent, and Crandell (2001) compared fourth-grade, fifth-grade, and principals' judgments of whether a given student read well enough to be promoted to the fifth grade. Overall, educators agreed with each other at about the 85% level. Educators agreed with the test results at about the 82% level when an original (lower) cut score was used for designating passing, and at about the 71% level when a more rigorous (higher) cut score was used. (Moving to the higher score had been part of the state's effort to adopt higher standards.)

Some tried to turn around the logic of using teacher judgment in standards-based reform, claiming not that teacher judgment was a good source of information for large-scale accountability, but that participation in scoring for accountability purposes would improve teachers' classroom instruction and assessment. As Goldberg and Rosewell noted (2000, p. 257), "Proponents of performance assessment often cite its [teacher participation in scoring] potential to drive instructional reform." Goldberg and Roswell (2000) tested that claim in the context of the Maryland State Performance Assessment Program (MSPAP). Scoring was done in the summers by 650 to 700 certified teachers. With samples of teachers with and without scoring experience, they used questionnaires, interviews, classroom visits, and classroom artifacts to investigate teachers' perspectives on MSPAP and its effects on classroom instruction and assessment. They were able to document a variety of good effects on instruction and assessment, including aligning activities to the Maryland Learning Outcomes, providing students more opportunities to read and write purposefully, and defining and using evaluative criteria. At the same time, there was some resistance to the amount of work involved and some surface learning. For example, teachers commonly confused performance assessments (the tasks themselves) with the learning they were supposed to indicate, conceiving of the task as the "learning" itself. They also often crafted or used rubrics that counted surface-level things or did

not set them in the context of larger performance objectives. The authors concluded (p. 289) that judgment-based scoring, by itself, yielded limited benefits and did not provide teachers with a "well-grounded understanding of performance-based instruction."

None of the state assessment systems described in this section are in place today. Current state assessment systems have less state-to-state variation, less performance assessment, and fewer uses of teacher judgment today, after the No Child Left Behind (2002) legislation, compared with the standards-based reform movement of the 1980s and 1990s.

"A few Bright Spots"

Teacher judgment used for scoring of writing assessment for state accountability purposes. Writing is the one "performance assessment" that has remained a part of most US state accountability systems, even after 2002. Writing assessment has face validity for the public, construct validity for an important educational aim, and teacher judgment reliability.

Acknowledging the problems the Vermont portfolio system was having with rater (teacher scorer) reliability, LeMahieu, Gitomer, and Eresh (1995) pointed out that both variability in teacher judgment and variability in the collection of student work contribute to unreliability of portfolio judgment. They demonstrated for writing portfolios in the city of Pittsburgh, Pennsylvania that with a scoring rubric developed with teacher input, and with attention to developing a shared understanding of the criteria within a coherent system of instruction and assessment, an acceptable level of reliability could be achieved.

In 1990, the state of Kentucky began work on a reform of both its assessment and accountability system and of schools and instruction more generally. Part of this reform effort included a writing portfolio, which did achieve acceptable scoring reliability (Fairtest, 1996). The Kentucky writing portfolio is still in use today. The program manages the accuracy of teacher (rater) judgment with conventional rater training procedures (KDOE, 2007).

Many other US states currently use on-demand writing assessments as part of their statewide accountability system, usually in conjunction with more conventional tests in reading, mathematics, and science. These state writing assessment programs usually use teachers as scorers, and they approach the issue of accuracy of rater judgment with conventional rater training procedures. Many of these states place summaries of their scoring procedures on the internet.

Teacher judgment of young children's work. Meisels and his colleagues developed the Work Sampling System (WSS; Meisels, Bickel, Nicholson, Xue, & Atkins-Burnett, 2001). This curriculum-embedded system used teacher judgments about young children's learning (age 3 to grade 5) as documented via checklists, portfolios, and summary reports. There are seven domains of development: personal and social, language and literacy, mathematical thinking, scientific thinking, social studies, the arts, and physical development. The system is very structured regarding both selection of student work and scoring on the domains, and has been found to be reliable when used by teachers. Meisels, Bickel, Nicholson, Xue, and Atkins-Burnett (2001) investigated the accuracy of teacher judgments of language and literacy and

mathematical thinking, using the WSS system, for reporting the development of students in Kindergarten through grade 3, with positive results.

"The one that Got Away"

Although the No Child Left Behind legislation effectively cut off most states' experimentation with teacher judgment or other participation in large-scale school accountability assessment except in writing, there was one notable exception. The Nebraska School-based, Teacher-led Assessment and Reporting System (STARS) program was a unique state accountability system, begun in 2000, toward the end of the standards-base reform movement and before NCLB, and dismantled in 2008.

Nebraska maintained a "conscientious objection" stance over against state testing longer than many expected would be possible.[1] The vision for STARS came from then Commissioner of Education Douglas Christensen, who believed strongly (Christensen, 1992) that

> Assessment must be as close to students as anything we do in the teaching-learning process. Assessment must help us get information about who our students are, what they learn, how they learn best, what they are able to do, what they know, and what they understand. There is no test that comes from outside the classroom that can do all of that.

Christensen (2000) further explained his vision as follows. Ironically, his argument relied in part on valuing teacher judgment – rhetorically expressed in the speech below in the negative, expecting the audience to find negative implications about teacher judgment unacceptable.

> Accountability cannot be so narrowly defined as reporting test scores. True accountability includes both what we do (processes and practices) and what results we get. Public reporting should require both….The assumptions behind school and classroom-based assessment and state testing are polar opposites. If you believe that state mandated standards and state created and mandated testing are good things, then you must accept two presumptions. One is the presumption of prevailing incompetence on the part of educators, and two is the presumption about the prevailing incapacity of educators and schools to set standards, design curriculum, deliver instruction and assess learning outcomes. Neither of these presumptions is acceptable to me.

As history showed, "state mandated standards and state created and mandated testing" won out, at least at the present time. But for at least eight years, Nebraska was known nationally for its position that local districts should be responsible for the assessment systems that monitor the progress of the students they teach (Roschewski, Gallagher, & Isernhagen, 2001).

The design of STARS. In STARS, school districts identified how they would measure and report student performance on content standards in grades 4, 8 and 11. They were allowed to select norm-referenced tests, develop criterion-referenced assessments, or use classroom assessments to measure state or state-approved content standards in reading and mathematics

[1] The author was a member of the National Advisory Council for the Nebraska STARS program from 2001 to 2008. She believes, however, this characterization of "conscientious objection" is not merely an opinion but is clearly supported by the literature cited here.

(Roschewski, 2004, p. 10). A statewide writing assessment, with student writing scored by trained teacher raters, was also in place (and still is).

District reporting of student achievement was accomplished by locally-developed assessment systems. The quality of district assessment *systems* was evaluated annually, beginning in 2001 (Buckendahl, Plake, & Impara, 2004). Six quality criteria (Plake, Impara, & Buckendahl, 2004) were used to provide technical quality ratings for district assessment systems:

1. Alignment of the test to the content standards
2. Opportunity for students to be exposed to the content prior to testing
3. Appropriateness of the assessments to the student population
4. Freedom from bias and sensitive situations
5. Consistency in scoring
6. Appropriateness of mastery levels

Districts submitted portfolios documenting how their assessment systems met these criteria. Thus, for example, to be rated "excellent" on Criterion 1, districts had to document that a panel independent of those who wrote the assessments had examined alignment of assessments to content standards, with what results. As a participant in the process, this author can attest that providing the results proved to be more difficult for districts than documenting panel meetings. Eventually, though, most districts (>99%, Schraw, 2007) did produce assessment portfolios showing their assessment systems met all these criteria. STARS received full federal approval in May of 2004, after two rounds of peer review and state response.

The quality of local assessments and the quality of teacher judgment. Several aspects of the STARS program did not entirely match the intended requirements of NCLB at first. Two issues were the quality of the local assessments themselves and the comparability of proficiency ratings from district to district, based as they were on performance on these different assessments rated by different teachers (Bandalos, 2004).

The State commissioned this author to do a study of the quality of the assessments themselves (Brookhart, 2004, 2005). The raters in the study, Nebraska educators invited by the state, reached a criterion of agreement during rater training. However, during the main scoring session the mathematics ratings remained reliable, while the reading ratings did not. Therefore, the published results (Brookhart, 2005) presented only quality ratings of local district mathematics assessments, which were high in most areas but medium to low in the reliability of scoring. The reading assessments, despite lack of reliability of ratings, displayed the same general pattern of quality ratings as the mathematics assessments (Brookhart, 2004). Using the rubric developed for that initial study, the state continued to regularly monitor the quality of assessments beginning in 2006 (Schraw, 2007).

The most important outcome of the STARS program, from many points of view (Gallagher, 2007; Isernhagen & Mills, 2007; Lukin, Bandalos, Eckhout, & Mickelson, 2004), was the increase in teacher assessment literacy and in teacher involvement in and ownership of the assessment process throughout the state of Nebraska as a result of STARS. Thus it is likely that the quality of teachers' judgments of student achievement became better during the STARS years. But the issue of comparability of judgment from district to district was not solved.

In the end, lack of comparability of proficiency ratings from district to district on the different local assessment systems – meaning achievement levels in Nebraska school districts could not be rank-ordered – caused the demise of the STARS program. Nebraska Legislative Bill 1157, approved by the governor on April 10, 2008, mandated the use of standardized tests to measure student achievement of state-adopted standards and charged the state board to "Provide for the comparison among Nebraska public schools and the comparison of Nebraska public schools to public schools elsewhere."

"Be Careful What You Wish For"

A potential – but as yet unproven – way forward for teacher judgment in summative assessment in the US may come through the standards-based grading movement. Teachers are beginning to see the inconsistencies between conventional grading practices and the student learning for which they are accountable. Many districts are changing their conventional report cards that typically listed subject areas and graded with the letters A, B, C, D, F or with percentages (0 through 100). The new standards-based report cards list standards (e.g., under mathematics might be listed "fractions, decimals, and mixed numbers," "concept of area," "problem-solving," and so on) and are typically graded with performance categories like Advanced, Proficient, Basic, and Below Basic.

This author knows of only three empirical studies related to standards-based grading to date. Two (Guskey, Swan & Jung, 2010; McMunn, Schenck & McColskey, 2003) report on initial efforts at employing standards-based grading after professional development, including perceived effects and difficulties. Only one study (D'Agostino & Welsh, 2007; Welsh & D'Agostino, 2009) spoke to the quality of teacher judgments in assigning standards-based grades.

D'Agostino and Welsh (2007) investigated the question of whether achievement-only, standards-based grading yielded accurate information about students in terms of their performance on state tests. Overall agreement rate between 3rd and 5th graders' standards-based grades in the district they studied and their proficiency levels on the Arizona Instrument to Measure Standards (AIMS, the Arizona State test) were only moderate: 44% for Mathematics, 53% for Reading, and 51% for Writing. When corrected for chance agreement, those percentages fell even lower (16% to 26%). In this regard, the authors corroborated Hoge and Coladarci's (1989) conclusions about the match between teacher judgment and standardized test scores.

However, D'Agostino and Welsh went on to investigate variation in the quality of teachers' judgment and its relationship to the quality of teachers' use of standards-based grading practices. In some classrooms, the convergence between graded and tested standards-based performance was much greater than in others. D'Agostino and Welsh (2007) hypothesized that teachers who used high-quality grading procedures (good judgment) would be more accurate in their appraisal of students against state standards than those who did not.

The researchers coded information from teacher interviews according to whether the following recommended grading practices were "clearly evident," "somewhat evident," or "not evident" (Welsh & D'Agostino, 2009, p. 85) in that teacher's grading:

- Assessing most of the performance objectives in state standards
- Grading on achievement, not effort
- Creating or obtaining assessments focused on state standards
- Identifying the objectives assessed by each assessment and tracking students' performance skill by skill
- Focusing on attainment of standards, not objectives listed in textbooks
- Using end-of-unit assessments for grading, not practice work
- Focusing on achievement, not progress (improvement)
- Assessing frequently
- Using multiple assessment approaches to measure different aspects of a skill
- Using a clear method for converting individual assessment results to standards-based grades

This list of attributes formed (with Rasch analysis) a unidimensional "Appraisal Style" scale. The two heaviest contributors to the scale were (a) being performance-focused, basing grades on standards of achievement rather than effort, and (b) assessing the full range of the performance objectives under a standard, not just some of them. The higher the Appraisal Style score, the more closely that teacher's grades agreed with his or her students' tested proficiency levels.

Thus this one study found that the higher the quality of teacher judgment in grading, the higher the quality of the information contained in the grades. While this is promising, research on standards-based grading is only in its infancy.

"Weighed in the Balance and Found Wanting"

A definitive conclusion is hard to reach. The results of one hundred years of study of teacher judgment in the US have been at best mixed. In both classroom level summative assessment (grading) and large scale summative assessment, teacher judgment has been found to be variable. For important public accountability functions, standardized tests are currently trusted over teacher judgment. One important exception is in the assessment of writing. However, standard rater training methods mold teachers into trained raters whose judgment matches predetermined examples. Rater training in the US does not use methods that take advantage of professional expertise, for example group moderation using professional conversations. The standards-based grading movement may help improve the quality of teacher judgment and the public impression of that quality, but it is too soon to tell for sure. Current glimmers of trust in teacher judgment have a long history of distrust to overcome.

References

Bandalos, D. L. (2004). Can a teacher-led state assessment system work? *Educational Measurement: Issues and Practice, 23*(2), 33-40.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals, Handbook I: Cognitive domain*. White Plains, NY: Longman.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill.

Brookhart, S. M. (1994). Teachers' grading: Practice and theory. *Applied Measurement in Education, 7*(4), 279-301.

Brookhart, S. M. (1994). Teachers' grading: Practice and theory. *Applied Measurement in Education, 7*, 279–301.

Brookhart, S. M. (2004, December). *Local district assessments: One indicator for the validation of Nebraska's standards, assessment, and accountability system*. Final report prepared for the Nebraska Department of Education.

Brookhart, S. M. (2005). The quality of local district assessments used in Nebraska's School-based Teacher-led Assessment and Reporting System (STARS). *Educational Measurement: Issues and Practice, 24*(2), 14-21.

Brookhart, S. M. (forthcoming). Grading and reporting. In J. H. McMillan (Ed.), *Handbook of research on classroom assessment.* SAGE.

Center on Education Policy. (2010, December). *State high school tests: Exit exams and other assessments*. Washington, DC: Center on Education Policy. Available: http://www.cep-dc.org/

Christensen, D. (1992, October 5). *Authentic assessment for the practitioner*. Speech delivered at the Nebraska ASCD Conference, Lincoln, NE.

Christensen, D. (2000, February 1). Keynote speech at the Nebraska Assessment Leadership Conference, Omaha, NE.

Cizek, G. J., Hirsch, T., Trent, E. R., & Crandell, J. (2001). A preliminary investigation of pupil proficiency testing and state education reform initiatives. *Educational Assessment, 7*(4), 283-302.

Coffman, W. E. (1971). Essay examinations. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 271-302). Washington, DC: American Council on Education.

Cross, L. H., & Frary, R. B. (1999). Hodgepodge grading: Endorsed by students and teachers alike. *Applied Measurement in Education, 12*(1), 53-72.

Cuban, L. (1993). *How teachers taught: Constancy and change in American classrooms 1890–1990* (2nd ed.). New York: Teachers College Press.

D'Agostino, J., & Welsh, M. (2007). *Standards-based progress reports and standards-based assessment score convergence.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Fairtest (1996, October). Kentucky portfolio scoring improves. Retrieved 4/22/11 from http://fairtest.org/kentucky-portfolio-scoring-improves

Forgione, P. D., & Doorey, N. A. (2010, December). New assessments for the common core standards, *NCME Newsletter, 18*(4), 7-17.

Fremer, J., & Dwyer, C. A. (1977, May). *Setting standards for basic skills in reading assessment*. Paper presented at the annual meeting of the International Reading Association, Miami, Florida. ERIC Document No. ED148833

Gallagher, C. W. (2007). *Reclaiming assessment*. Portsmouth, NH: Heinemann.

Gamson, D. (2007). Historical perspectives on democratic decision making in education: Paradigms, paradoxes, and promises. In P. Moss (Ed.), *Evidence and decision making* (pp. 15-45). The 106th yearbook of the National Society for the Study of Education, Part I. Malden, MA: Blackwell.

Goldberg, G. L., & Roswell, B. S. (2000). From perception to practice: The impact of teachers' scoring experience on performance-based instruction and classroom assessment. *Educational Assessment, 6*(4), 257-290.

Guskey, T. R., Swan, G. M., & Jung, L. A. (2010, May). *Developing a statewide, standards-based student report card: A review of the Kentucky initiative.* Paper presented at the annual meeting of the American Educational Research Association, Denver. ERIC Document No. ED509404

Haertel, E., & Herman, J. (2005, June). *A historical perspective on validity arguments for accountability testing.* National Center for Research on Evaluation, Standards, and Student Testing (CRESST),University of California, Los Angeles. ERIC Document No. ED488709.

Harlen, W. (2005). Trusting teachers' judgment: research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education, 20*(3), 245-270.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59*(3), 297-313.

Ingersoll, R., & Merrill, L. (2010, Fall). The changing face of the teaching force. *@Penn GSE: A Review of Research.* Retrieved 4/20/11 from http://www.gse.upenn.edu/review/feature/ingersoll

Isernhagen, J. C., & Mills, S. J. (2007). *Charting STARS: Engaging Conversations (STARS Year Seven Evaluation).* Nebraska Department of Education and University of Nebraska-Lincoln. Available: http://www.education.ne.gov/assessment/STARSEvaluation.htm

Kentucky Department of Education (2007). *Kentucky writing handbook: Part II: Scoring*. Retrieved 4/22/11 from http://www.education.ky.gov/users/jwyatt/writing/KyWritingHandbook/Grade%2012%20Scoring/Complete%20Scoring%20Handbook%20Grade%2012.pdf

Koretz, D., Stecher, B., Klein, S.., & McCaffrey, D. (1994). The Vermont Portfolio Assessment Program: Findings and implications. *Educational Measurement: Issues and Practice, 13*(3), 3-16.

LeMahieu, P. G., Gitomer, D. H., & Eresh, J. T. (1995). Portfolios in large-scale assessment: Difficult but not impossible. *Educational Measurement: Issues and Practice, 14*(3), 11-28.

Lortie, D. (1975). *School Teacher*. Chicago: University of Chicago Press.

Lukin, L. E., Bandalos, D. L., Eckhout, T. J., & Mickelson, K. (2004). Facilitating the development of assessment literacy. *Educational Measurement: Issues and Practice, 23*(2), 26-32.

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice, 20*(1), 20-32.

McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *Journal of Educational Research, 95*, 203-213.

McMunn, N. Schenck, P., & McColskey, W. (2003, April). *Standards-based assessment, grading and reporting in classrooms: Can district training and support change teacher practice?* Paper presented at the annual meeting of the American Educational Research Association, Chicago. ERIC Document No. ED475763

National Commission on Excellence in Education. (1983). *A nation at risk*: *The imperative for educational reform.* Washington, DC: US Government Printing Office. Available: http://www2.ed.gov/pubs/NatAtRisk/index.html

Nebraska Legislative Bill 1157 (2008). *Change provisions relating to the statewide system for assessment and reporting of student learning***.** Retrieved 4/22/11 from http://www.nebraskalegislature.gov/FloorDocs/100/PDF/Slip/LB1157.pdf

No Child Left Behind Act of 2001. Pub. L. No. 107–110, 115 Stat. 1425 (2002).

Plake, B. S., Impara, J. C., & Buckendahl, C. W. (2004). Technical quality criteria for evaluating district assessment portfolios used in the Nebraska STARS. *Educational Measurement: Issues and Practice, 23*(2), 12-16.

Porter, A. C. (1993). School delivery standards. *Educational Researcher, 22*(5), 24-30.

Randall, J., & Engelhard, G. (2010). Examining the grading practices of teachers. *Teaching and Teacher Education, 26*(7), 1372-1380.

Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal, 48*(2), 335-360.

Roschewski, P. (2004). History and background of Nebraska's School-based Teacher-led Assessment and Reporting System (STARS). *Educational Measurement: Issues and Practice, 23*(2), 9-11.

Roschewski, P., Gallagher, C., & Isernhagen, J. (2001). Nebraskans reach for the STARS. *Phi Delta Kappan, 82*, 611-615.

Schraw, G. (2007, November). *Report on the Nebraska 2006-2007 peer review process*. Nebraska Library Commission. Retrieved 4/22/11 from http://www.nlc.state.ne.us/epubs/E2420/B024-2007.pdf

S.G.B. (1840). Weekly reports in schools. *The Common School Journal, 2,* 185–187. Reprinted in Laska, J. A., & Juarez, T. (1992). *Grading and marking in American schools: Two centuries of debate* (pp. 11–14). Springfield, IL: Charles C. Thomas.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22-28.

Starch, D., & Elliott, E. C. (1912). Reliability of the grading of high-school work in English. *School Review, 20,* 442–457.

Starch, D., & Elliott, E. C. (1913a). Reliability of grading work in mathematics. *School Review, 21,* 254–259.

Starch, D., & Elliott, E. C. (1913b). Reliability of grading work in history. *School Review, 21,* 676–681.

Stiggins, R. J., Frisbie, D. A., & Griswold, P. A. (1989). Inside high school grading practices: Building a research agenda. *Educational Measurement: Issues and Practice, 8*(2), 5–14.

Tyack, D. (1974). *The one best system*. Cambridge, MA: Harvard University Press.

Welsh, M. E., & D'Agostino, J. V. (2009). Fostering consistency between standards-based grades and large-scale assessment results. In T. R. Guskey (Ed.), *Practical solutions for serious problems in standards-based grading* (pp. 75-104). Thousand Oaks, CA: Corwin Press.