Academic Leadership: The Online Journal

Volume 6
Issue 2 *Spring 2008*

Article 16

4-1-2008

Reliable Sources: Recruiting and Developing Evaluators, External to the University Community

Paul Watkins

Ruth Roberts

Academic Leadership Journal

Wiggins and McTighe (2005, 18) challenge educators to think critically about acceptable assessment evidence by asking: “How will we know if students have achieved the desired results? What will we accept as evidence of student understanding and proficiency?” (p 18). Teacher education programs must face these important questions and affirm that answers are both valid and equitable. This article explores the benefits of evaluator training in the scoring of high-stakes work samplings produced by teacher preparation candidates.

At Southeast Missouri State University, completion of a Teacher Work Sample (TWS) provides authentic evidence that a teacher candidate has the ability to design learning strategies and assessment goals that enhance student learning. The Teacher Work Sample further demonstrates a candidate’s skill and competence to become a certificated teacher in the state of Missouri. As a result, the successful completion of the TWS holds a high priority with faculty and teacher preparation candidates. Further, the student work sampling is a priority assessment that meets both national and state standards for teacher preparation programs. It is also a critical component for National Council for Accreditation of Teacher Education (NCATE) (Mitchell, Allen, and Ehrenburg 2006), Missouri, and the university’s on-going teacher preparation assessment program.

NCATE requires that teacher preparation programs use practical assessment and that they are systematically evaluated to validate their value (NCATE 2003). This validity is evaluated on several levels of measurement: 1) Meaningful knowledge and skills taught; 2) Fairness; and 3) Robustness of results. All three of these measures address the credibility of the assessment. The Missouri Department of Elementary and Secondary Education (DESE) also acknowledges the value of the TWS as a means of meeting the eleven quality indicators teacher candidates must meet for Missouri teacher certification (DESE 2006). Both NCATE and DESE tacitly expect that assessments, such as the TWS, are scored reliably; using a clear set of expectations shared with evaluators and teacher candidates.

The Teacher Work Sample at Southeast Missouri State University is arranged in two parts, writing prompts (including standards) and scoring rubric. Writing prompts define the eight elements the teacher candidates must address (1) contextual factors, (2) learning goals, (3) assessment planning, (4) designing instruction, (5) analyzing assessment, (6) involving families, (7) managing the classroom, and (8) reflecting and self-evaluating professional practice (Teacher Work Sample Committee 2006). The scoring rubrics assess three levels of competence for each of the eight elements. For instance, an element can have no credible evidence and “Not Meet” the standard. An element may only “Minimally Meet” a standard, demonstrating irrelevant or partial knowledge of an element. Or, candidates may show that they “Partially Meet” the element’s expectation. However, for candidates to demonstrate total command of the skill, they must fully “Meet” the standard. The writing prompts and rubrics represent a critical framework for assessing the level of teaching performance bench marked during each of the three semesters in the teacher preparation program.

Candidates in their final bench-mark semester, student teaching, must construct a summative TWS,

which is submitted online and evaluated online. They must “Meet” all of the element criteria scored by the rubric before applying for a Missouri certification or completing the teacher education program. As a result, a successful TWS becomes a high stakes enterprise for these candidates and requires a fair and meaningful rating by university faculty and others who score it.

Recruiting and Training External Evaluators

Because the work sampling provides a vital gateway to certification, the university’s college of education expects quality assurance that scoring results reflect valid and reliable results. Only a limited number of university faculty are available to score Teacher Work Samples; consequently, the college looks outside the University for qualified external scorers to help relieve the scoring magnitude faculty face. In the fall 2006, fourteen teachers from surrounding schools were recruited to help score the TWS. These teachers represented schools that regularly host student teachers for the university and teachers, who themselves have worked with student teachers. Their willingness to assist with scoring connects the teacher preparation program to the PreK-12 student progress and learning.

These classroom teachers were screened through the recommendations of the student teacher supervisors who worked with them as cooperating teachers. Supervisors looked for teachers who understood the teacher preparation program and believed in its mission. They had to be familiar with the university requirements for successful completion of a sixteen-week field experience. Faculty recommendations were another vital source of screening. They worked closely with PreK-12 teachers over an extended period in various field assignments through early preparation classes.

Building principals also provided valuable insight into the professional practice of scoring participants. They observed first-hand these Pre-K-12 teachers as they worked with student teachers in the classroom. They knew these teachers were professionals who could be counted on to complete a task and who could mentor colleagues in meaningful ways. With the triangulation of references (supervisor, faculty, and principal) the university was confident that a pool of qualified external scorers had been identified. Training them as qualified and impartial scorers remained.

Getting Started

Once the screening process discovered qualified external raters, they were invited to participate as part of the scoring team. As a condition for scoring, participants were required to attend an evening institute for TWS evaluation. They were provided the

Teacher Work Sample Handbook with scoring standards, prompts, and rubrics as a framework for their training. The external raters were first introduced to the guiding philosophy behind the TWS as the teacher education capstone evaluation. This philosophy believes that the TWS builds a habit of mind for planning instruction and assessing PreK-12 student learning. It further promotes teaching as a reflective practice, which results in continuous professional improvement.

Training required participants to first look carefully at the construction prompts and criteria students use to develop their work sample. The Teacher Work Sample organizes the teacher candidate responses around each of the eight elements. Candidates construct their written evidence by following a series of element prompts. For instance, the first element in the TWS asks students to examine carefully the context of teaching and learning. The standard (Teacher Work Sample Committee 2006, for this

element states that, “The [teacher candidate] uses information about the learning-teaching context and state individual differences to set learning goals and plan instruction and assessment”. This standard is followed by a series of prompts that include (1) community, district and school factors (2) classroom factors (3) student characteristics, and (4) instructional implications. These standards and prompts provide important background for raters to know. Raters had to understand clearly the relations among the standard and performance expectation for each element. Raters also became aware of how the rubric captures each dimension of performance that must meet the standard before they could begin reading work samples with clarity and focus.

Next, raters made a quick read of the TWS and scored it considering only their initial reaction without the influence of anyone else. Next, participants were asked to score the model again; this time as a collaborator in groups of two, discussing weaknesses and strengths of the model. They were encouraged to find common ground for their final assessment and agree within 10 points on a representative score. This collaborative score was then compared to earlier scoring results. Participants, as a result, saw a more reliable evaluation evolving as a result of their collaborative efforts. Talking through the rubric and examining the model response together was a much more informed approach to scoring than a single scorer’s independent result.

Once the performance standards were firmly set in the minds of the raters, they more clearly discriminated between a criterion that meets and one that does not. These levels of success were outlined earlier. Scoring points assigned to each level of evidence calibrates the degree of quality achieved for each criterion. Participants found that it was important to discuss and clarify with each other the interpretations of these levels and agree on a common rating.

External raters, particularly, must have a firm understanding of the TWS scoring model because they do not teach the TWS in the various techniques classes; as a result, they lack a thorough background for the TWS and where additional scoring evidence might be found in work outside the prescribed element or criteria (Denner et al. 2003). Table 1 is a matrix used to help raters find alternative elements in the TWS where additional supporting evidence is imbedded.

Table 1.

Finding Evidence for the TWS: Southeast Missouri State University

TWS Elements/ Teaching Standards	Contextual Factors	Learning Goals	Assessment Plan	Design for Instruction	Classroom Management
The Candidate uses information about the learning/teaching context and student individual differences to	X	X	X		

set learning goals, plan instruction and assess learning.						
The Candidate sets significant, challenging, varied and appropriate learning goals.	X	X		X	X	
The Candidate uses multiple assessment modes and approaches aligned with learning goals to assess student learning before, during and after instruction.			X		X	
The Candidate designs instruction for specific learning goals, student characteristics and needs, and learning contexts.	X	X		X	X	
The Candidate uses an understanding of individual and group motivation and behavior to create a learning environment that encourages social interaction, active engagement in learning and self-motivation.	X				X	X

The Candidate uses on-going analysis of student learning to make instructional decisions.	X	X	X	X	X	
The Candidate uses assessment data to profile student learning and communicate information about student progress and achievement.	X	X	X			

Table 1

Finding Evidence (Continued)

TWS Elements/ Teaching Standards	Contextual Factors	Learning Goals	Assessment Plan	Design for Instruction	Classroom Management	Instru Dec Ma
The Candidate analyzes the relationship between his or her instruction and student learning in order to improve teaching practice	X	X	X	X	X	
The Candidate	X	X		X		

involves children's families in the unit of study. A strong home-school connection is important for the children's success in this unit.						

X = Primary Evidence

X = Secondary Evidence

The table above proved valuable for scorers. The cross references to evidence helped them see the TWS as a complete document, not one that is compartmentalize into eight distinct elements.

Scoring Bias

A complete understanding of the TWS elements and criteria can also protect scoring from teacher bias. Bias, like a hidden predator, must be revealed (Gay 1996) or it will taint an otherwise effective Teacher Work Sample score. It was important at the training that scoring biases were part of the discussion during training. Biases such as: how a classroom is best managed, cooperative learning as the best teaching strategy, or poor spelling and grammar as reflecting poor teacher practice. These were only a few bias triggers that were challenged by the trainer. These biases could unfairly impact an otherwise competent TWS (Szpara & Wylei 2005). While it can never be assured that bias is eliminated from scoring, the training session did help address those issues that lay just below the surface for scorers.

During their training, scores developed awareness that the university students are novice practitioners who do not have the high level teaching skills of a veteran. One confounding bias among practitioners who work with student teachers is that they are ready to take over and be the professional. If this were true an internship would not be necessary. As a result, a good deal of discussion was devoted to skills a novice student teacher should be expected to bring to the classroom. These skills included competencies they would want to see such as dispositional attitudes of warmth, openness with their students, and consistency with expectations. They also listed more pragmatic skills such as competency in subject content, communicating clear expectations of students, and willingness to take criticism. Thus, a TWS document they score may not represent the level of teaching competency that might be expected from a teacher with three or more years of experience. Yet it should be expected to reveal the basic skills scorers agreed a novice should bring to their student teaching experience. As the scorers approach their task of evaluation they must be sensitive to that reality.

Simply training external scores would not be enough assurance that results were reliable. The next step required a reliability study confirming the scoring was fair and consistent.

Quality Scoring Assurance

Raters were placed into scoring teams of two raters and assigned five or six Teacher Work Samples to score. Results were to be submitted to students within seven days of the team's assignment. All of the teams successfully scored their Teacher Work Samples within the seven-day deadline. Once published to the teacher candidates, the concern remained: how closely did the teams score their work samples? Did training make a difference in the effectiveness of the external scorers?

Fourteen external scorers participated in the fall 2006 scoring of the student teaching Teacher Work Sample. The participants worked as classroom practitioners in local districts that host Southeast teacher candidates in their final semester of teacher preparation. The average scorer's teaching experience was 13 years. Ten scorers were elementary teachers teaching in grades ranging from kindergarten to grade six. Two were secondary teachers, in the content areas of social studies and English. One scorer was a special education teacher. The scorers were highly motivated to do the work and bring quality to their results. They were assisted in their efforts with online services which allowed flexibility in scoring times and electronic collaboration.

Teacher candidates uploaded their Teacher Work Samples to secure web site. Once these were uploaded, external raters were then provided access to the work samples entrusted to them for scoring. The web site used for scoring allowed raters to collaborate through online messaging or they could use e-mail. They did not have to sit in the same room or the same building with their colleague looking at a common screen.

The collaboration time online was more flexible because the scoring availability was any time, day or night. Certainly, the convenience of online scoring was an advantage, but it also provided a feedback mechanism for scorers. Clarification and immediate questions did not have to wait for returned phone calls or postage to be delivered. The TWS administrator also had the advantage of electronically monitoring the scoring progress toward assigned deadlines. Having this sophisticated feedback capability (Reid & DeMaster 1972) proved critical to the scoring validity among the external scorers.

Confidence in the TWS scoring results provided by scorers online and outside the university drove the purpose of this study. With training, a clear statement of scoring expectations, and years of teaching experience it was believed that, yes, external scorers could reliably evaluate work samples.

Scores from the fourteen external scorers were downloaded from the website data base and analyzed. A Pearson Correlation of paired cores was calculated (Table 2).

Table 2.

Pearson Correlations for External Scorers of the Teacher Work Sample

	First Score	Second Score
First Score	Pearson Correlation	1	.903**

Sig. (2-tailed) .000

N* 37 37

Second Score Pearson Correlation .903** 1

Sig. (2-tailed) .000

N* 37 37

* N = Number of TWSs scored

**Correlation is significant at the 0.01 level (2-tailed)

A general interpretation of a correlation coefficient between .8 and 1.0 is considered very high (Kerlinger 1973). Thus, a correlation of .903 demonstrates a meaningfully high correlation. The corollary results indicate that there is a strong scoring reliability between external scorers (Table 2) first and second score. This evidence helps confirm that training was a strong contributor to scoring reliability. However, a clear scoring guide and teaching experience cannot be ruled out as contributing variables.

The correlations resulting from this study are encouraging. Relying on external raters to assist faculty in scoring the final Teacher Work Samples does not diminish the reliability of the scores. The results also confirm that training external scorers for the scoring is critical in maintaining reliable TWS scores. These results are encouraging for the future training of external scorers and the addition of classroom teachers to the surge of expert scorers.

Conclusion

External scorers provided a larger pool of labor and lifted much of the scoring load from university faculty, but discovered personal benefits as well. They learned that teacher candidates must meet high standards and demonstrate their influence on student learning. The TWS scoring also impacts the professional growth of external scorers. In the words of one external scorer, “I truly appreciated the scoring opportunity. It provided a unique insight into the expectations of the university’s teacher preparation program.”

The results of this study have helped the College of Education answer Wiggins and McTigue’s earlier questions, “What will we accept as evidence of student understanding and proficiency?” Also, it brought practitioners and future practitioners into a venue where learning for both is a result of shared information and ideas. Information and ideas which can be interpreted and challenged resulting in the reflection and rethinking that promotes quality professional growth.

References

Denner, P. R., A. D. Norman, S. Salzman, R. S. Pnakratz, and C. S. Evans. 2004. The renaissance partnership teacher work sample: Evidence supporting score generalizability, validity, and quality of student learning assessment, 23-55. Dubuzue, IA: Kendall/Hunt Publishing Co.

Department of Elementary and Secondary Education. (2006).

Missouri

standards for teacher education programs (MoSTEP): Benchmarks for preliminary teacher education programs. Retrieved June 7, 2007, from http://www.dese.mo.gov/schoollaw/rulesregs/Inc_By_Ref_Mat/MoSTEP%20(finalrevisedversion-OrderRulmkng)%2010-06.pdf.

Gay, L. R. (1996).

Educational research: Competencies for analysis and application. Columbus OH: Printice-Hall, Inc.

Kerlinger, F. N. (1973).

Foundations of behavioral research, 2nd edition. New York: Holt, Rinehart and Winston.

Mitchell, A., A. Sheila, and P. Ehrenburg. 2006.

Spotlight on schools of education: Institutional responses to NCATE Standards 1 and 2. Washington D. C.: National Council for Accreditation of Teacher Education.

National Council for Accreditation of Teacher Education

. Assessing education candidate performance: A look at changing practices, ed. Elliott Emerson (Washington D. C.: National Council of Accreditation of Teacher Education, 2003).

Reid, J. B. & DeMaster, B. (1972).

The efficacy of the spot check procedure in maintaining the reliability of data collected by observation quasi natural settings: Two Pilot studies. Oregon Research Bulletin. 12(8) pp.

Renaissance Partnership Colleagues (2003, June).

“How to” Part of the renaissance partnership credibility manual. (Issue June 18, Work session). St. Louis: Renaissance Group.

Szpara, M. Y. (2005). National board for professional teaching standards assessor training: Impact of bias reduction exercises. Teachers College Record 107(4) pp. 803-841.

Teacher Work Sample Committee. (2006).

Teacher work sample: Student guidelines. Cape Girardeau: Southeast Missouri State University, College of Education.

Wiggins, G. & McTighe, J. (2005).

Understanding by design. Alexandria VA: Association of Supervision & Curriculum Development.