Vol. 2, No. 2, Feb 2014, pp. 3-18

The Renaissance Group

Examining the Reliability of a Culminating Teacher Education Assessment and Discovering Areas for Reform

Lisa D. Murley, Rebecca Stobaugh, Pamela Jukes, Janet Tassell
Western Kentucky University

Abstract

The purpose of this article is to provide an overview of the process used to examine the inter-rater reliability of the Teacher Work Sample (TWS) Scoring Rubric involved with the senior culminating experience for teacher candidates used at a large comprehensive university. The study compared holistic and analytic scores reported by Student Teacher Seminar course instructors to those of trained participants to determine the consistency of ratings between the two groups. The study resulted in several clear areas for revising the TWS for reliability and created a foundation for future revisions. What may prove to be the most important finding of the study, however, is the need to examine the differences among scoring practices of raters because scoring varies among people. Common errors include misinterpretation of scoring rubrics, prompts, the teaching and learning process, and even concepts such as revised Bloom's Taxonomy. This finding could be generalized to other universities as all education programs utilize scoring prompts and rubrics to measure teacher candidate performance and most all use revised Bloom's Taxonomy in the teaching and learning process.

The National Council for Accreditation of Teacher Education (NCATE 2000) required teacher preparation programs to document the success of teacher candidates with the intent of affecting a "…paradigm shift from a focus on the teaching process to learning results and connecting teacher performance to student learning" (Pankratz, 2000, p. 1). One instrument used by universities across the nation to measure teacher candidate ability to impact P-12 student learning and teacher candidate preparation and performance on teaching process is the Teacher Work Sample (TWS). Developed by the Renaissance Partnership, a consortium of 11 universities, the TWS as defined by Kohler, Henning, and Usma-Wilches (2008, p. 2109), is a "…performance-based assessment tool that enables teacher education programs to provide evidence of student teacher ability to meet state and national teaching standards" (as cited in Girod, 2002; McConney, Schalock, & Schalock, 1998; Schalock & Myton, 1988; The Renaissance Partnership for Improving Teaching Quality, 2001).

Many teacher education programs adopt performance-based assessments to measure candidate performance against professional standards that require reflection and evaluation of the candidate's own practice based upon a set of standards (Darling-Hammond & Bransford, 2005). The TWS engages candidates in this type of rigorous performance assessment by using state teacher standards as the criteria for the measurement of proficiency. The scoring rubric is used to measure teacher candidate performance in the following teaching processes believed to be important in the teaching and learning process (Tassell, Stobaugh, & McDonald, 2013):

- Contextual Factors: Identifying relevant contextual factors and exploring how these factors may affect the teaching-learning process.
- Learning Goal & Pre/Post Assessment: Creating and justifying Learning Goals for the unit; designing and pre/post assessment to monitor student progress toward learning goals.
- Design for Instruction: Analyzing student performance on pre-assessment; constructing a unit overview that addresses Learning Goals and student needs; describing instructional strategies and formative assessments that include a technology component.
- Analysis of Student Learning: Representing, analyzing and communicating assessment data for all students and significant subgroups; analyzing decisions made regarding the instruction and assessment to determine the success of instruction.
- Reflection: Reflecting on performance as a teacher and linking the performance to student learning results and state teaching standards; evaluating the performance and identifying future actions for improved practice and professional growth.

A scoring rubric is used to score the candidate's TWS. The TWS analytic scoring rubric is based on a 4-point scale (1= Beginning; 2= Developing; 3 = Proficient; 4 = Exemplary). In addition, each TWS is assigned a holistic score using the same scale (School of Teacher Education, 2011).

One large, comprehensive university was a charter member of the Renaissance TWS group committed to using TWS as a tool for instruction and as a performance assessment of teacher candidates. The university piloted the TWS in 2001 and beginning in 2003 required teacher candidates to submit a completed TWS for grading in a Student Teaching Seminar course scheduled for the last semester of the candidate's undergraduate teacher preparation program. The TWS requires the candidate to demonstrate the capacity to positively impact student learning as they plan, deliver, and assess a standards-based unit of instruction, analyze the results

of student assessments, and reflect on the effectiveness of instruction and student learning to improve instruction.

Due to the data analysis of teacher candidate scores to determine strengths and weakness of the TWS, the education faculty formed a TWS Task Force Committee in January 2010 charged with the task to revise the teacher work sample with a targeted implementation date of Fall 2010. The TWS Task Force was comprised of education and content area university faculty as well as P-12 teachers who supervise the teacher candidate student teaching experiences in the P-12 classrooms. After reviewing the strengths and weaknesses of the TWS, the Task Force decided upon the following goals as the primary focus for the work:

- Align the TWS more fully with the state teacher standards (Kentucky Department of Education, 2008).
- Improve the TWS components which include the prompt and scoring rubric.
- Clarify and streamline prompts and scoring rubrics.
- Increase rigor and promote higher levels of performance.
- Increase scoring consistency to promote reliability (Nitko & Brookhart, 2010; Popham, 2009).

Based on these goals, the TWS Task Force revised the TWS components, prompts, and scoring rubrics to reflect the needed improvements. The revised TWS was implemented into practice in the Fall 2010 academic term. Additional TWS data collection continued throughout that same term from students, faculty, and P-12 practitioners to determine any further changes that needed to be made. The TWS was revised and the full implementation began the Spring 2011 academic term. The Task Force met at the end of the semester after one year of full implementation to decide if revisions should be made. Examination of the recently collected teacher candidate data helped the Task Force determine that further exploration of the TWS prompt and scoring rubric was needed (Stobaugh, Tassell, & Norman, 2012).

Purpose and Research Questions

As a part of the revision process, the TWS Task Force focused on TWS scorers' inter-rater reliability of the TWS in an effort to glean insights related to the prompts, scoring rubric, and learning processes. Therefore, the purpose of the study was to examine the inter-rater reliability of the TWS Scoring Rubric after the implementation of the revised TWS in the Spring 2011 academic term.

The research questions for the inter-rater reliability study were:

1. How do university Student Teacher Seminar course instructors' TWS scores, both holistic and analytical, compare to scores of a mixed group of TWS trained university faculty and P-12 practitioners trained specifically for the TWS scoring session?
2. What do qualitative data collected from scoring participant comments reveal about the TWS scoring prompts and rubrics?

Method

Teacher Work Sample submissions were selected for examination from all those submitted by the teacher candidates during the Spring 2011 academic term. To ensure the samples represented the appropriate percentage of teacher candidates from each program area (e.g., elementary, middle, secondary), 100 samples were randomly selected across program areas to correlate with the enrollment of teacher candidates in each program area. Thus, for example, the largest number of samples came from Elementary Education because the program area has

the largest number of teacher candidates. Table 1 presents the number of student teachers in each program area in Spring 2011 and the number of TWS samples randomly selected within each program area from that semester.

Table 1
Distribution of Number of Samples by Program Area

Program Area	Semester Total (n=182)	Sample (n=100)
Early Elementary P-5	95	52
Middle Grades 5-8	25	14
Early Childhood P-K	8	4
Agriculture 5-12	2	1
Art P-12	6	3
Business 5-12	6	3
Family and Consumer Science 5-12	8	5
Music P-12	12	7
Physical Education 5-12	10	6
Spanish 5-12	2	1
English Secondary	6	3
History Secondary	2	1

Pre-Scoring Training

Prior to the scoring sessions, two samples were selected for quality control and TWS scoring training purposes. These samples were purposefully selected by the researchers to represent a "Proficient" and "Developing" typical TWS. The researcher requested the work samples from an instructor of the Student Teacher Seminar course. The instructor was asked to remove all identifiers such as the teacher candidate and school names and to send the work samples electronically to the researchers. Study participants who had not received TWS scoring training prior to the study, along with all other participants who had received the training through other scoring experiences, were required to score the two selected TWS submissions before the TWS training session and submit the ratings to the researchers electronically. The researchers compiled a comprehensive chart of the results depicting the holistic score along with the analytic score from each of the scoring rubric indicators.

Training Session

At the training session of the research study the participants discussed the pre-training results which included discussion of score deviations of three or more levels from one of the samples. Some biases were exposed such as one participant stated that since the TWS student product was better than previous semesters when she had scored, she determined this created a bias which had inflated the scores.

The trainers provided more clarification of the expectations of the TWS scoring rubric and addressed any questions related to scoring or scoring procedures. Participants had opportunity to become familiar with the levels for scoring listed at the top of the TWS rubric: Beginning, Developing, Proficient, and Exemplary.

Individual Scoring

After the training session, each participant received a CD with ten TWS and a score sheet. Participants individually scored each TWS sample and returned score sheets to the

researchers prior to the second meeting of participants so the scores could be compiled and analyzed. Participants were given a comments page to record concerns, questions, or feedback about each indicator when scoring the TWS.

Data Collection

Both quantitative and qualitative data were collected at the scoring session to answer the research questions. The data were comprised of the TWS scores as well as the qualitative data collected during the scoring session. Descriptive statistics were used to summarize the TWS scores and determine their differences. Study participant scores for each TWS holistic score and analytic scores given for the indicators on the scoring rubric were compared to the scores assigned by the Student Teacher Seminar course instructors for Spring 2011.

Qualitative data included participant comments about the scoring process and specific scoring rubric indicators. Participants submitted ideas on a comments page about each indicator after scoring the TWS. In addition, a follow-up discussion occurred in which scores and areas of disagreement were shared.

Data Analysis

The participant scores were compared to scores assigned by the Spring 2011 Student Teacher Seminar course instructors. Statistics for differences, including standard deviation per indicator, were determined to identify those indicators whose scores differed by more than two performance levels: 1 = Beginning, 2 = Developing, 3 = Proficient, and 4 = Exemplary. The frequency and percentages of differences of scores for each indicator were also determined.

Results

The results from the data collection are presented in Table 2. The standard deviation for the holistic score and each indicator are shown of the scored samples. Parts with the greatest standard deviations included Learning Goal and Pre/Post Assessment Plan indicators 2, 4, 6, and 9; Design for Instruction indicator 4; and Reflection of Teaching Practices indicator 3. Indicators with the smallest standard deviations included the first indicator in both the Learning Goal and Pre/Post Assessment and Reflection on Teaching Practices sections. Table 2 also provides the number of times scores differed by more than two score levels, that is, how many times participant scores and Student Teacher Seminar Course instructor scores had a difference of two or more performance levels on a given indicator.

The results suggest that holistic scores tend to be fairly consistent (SD=0.75) overall. Both university course instructors and trained project participants scored within one point of each other across the Teacher Work Samples. These results suggest that the TWS Scoring Rubric may be somewhat reliable when used holistically.

Table 2
Standard Deviation and Holistic Score Different by More Than Two Levels

Rubric Indicator	Standard Deviation	Score Different by More Than Two Levels
Holistic Score	0.75	11
Contextual Factors (CF)		
CF1 School Information	0.80	11
CF 2 Classroom Information	0.78	11
CF 3 Student Characteristics	0.85	10
Learning Goal Pre/Post Assessment (LG)		
LG 1 List 2 to 3 Learning Goals	0.67	5
LG 2 LG Levels	0.92	17
LG 3 LG Alignment to Standards	0.81	12
LG 4* LG Appropriateness	1.05	21
LG 5 LG Mastery Levels	0.71	6
LG 6* Pre-Post Assessment : LG	1.05	20
LG 7 Pre-Post Assessment: Modes	0.85	13
LG 8 Modes of Assessment	0.71	9
LG 9* Assessment Scoring Criteria	1.19	14
Design for Instruction (DFI)		
DFI 1 Results of Pre-Assessment	0.50	6
DFI 2 Unit Overview	0.80	9
DFI 3 Integration of Technology	0.86	11
DFI 4* Instructional Strategies	0.97	16
DFI 5 Formative Assessments	0.85	10
Analysis of Student Learning (ASL)		
ASL1 Visual Representations	0.90	6
ASL2 Performance Analysis	0.77	4
ASL 3 Instructional Implications	0.87	9
ASL 4 Individual Student	0.79	8
Reflection of Teaching (ROT)		
ROT 1 Self-Assessment	0.58	2
ROT 2 Teaching Strengths	0.77	6
ROT 3* Professional Development	0.98	15

*Largest standard deviations

These results suggest that data for five indicators surface as cause for concern. We focused on SD levels that neared or exceeded 1.0. Scores for five indicators tended to be less consistent (SD=1.04, 1.04, 1.19, 0.97, 0.98). A discussion follows of each of these indicators with the inclusion of the qualitative data collection from the scoring participant comments in an effort to delve deeper in the overall data collection of the research study.

Learning Goal and Pre/Post Assessment: Indicator 4 (LG 4)

The fourth indicator in the Learning Goal and Pre/Post Assessment section had a high standard deviation of 1.05. The indicator states, "Clear and logical justification in the 4 required areas for learning goal appropriateness: student prior knowledge, student learning needs and/or developmental appropriateness, authentic real world, and other relevant connections" (School of Teacher Education, 2011, p. 12). Table 4 displays the indicator cells from the TWS scoring rubric.

Table 4
TWS Scoring Rubric Indicator Learning Goal 4 (LG4)

Indicator	Beginning	Developing	Proficient	Exemplary
LGA 4 Appropriateness of Learning Goals	Justification is missing for two goals Or 2 or more justifications of the required areas in the prompt	Justification is missing for one goal Or 3 or more justifications of the required areas in the prompt	Clear and logical justification in the 4 required areas for learning goal appropriateness: student prior knowledge, student learning needs and/or developmental appropriateness, authentic real world, and other relevant connections.	Achieves the Proficient level with minimal assistance on the first attempt and demonstrates above and beyond the Proficient level.

For this indicator all four required items are necessary for a teacher candidate to obtain a proficient score: (a) student prior knowledge, (b) student learning needs and/or developmental appropriateness, (c) authentic real world, and (d) other relevant connections. Teacher candidates often included some of the four items, but not all of them. Scorers may inconsistently score this section because they fail to check for all four items. The Developing level states, "Justification is missing for one goal OR 3 or more justifications of the required areas in the prompt" (School of Teacher Education, 2011, p. 12). A participant scorer questioned which category to mark if a justification was inappropriate or inaccurate. Several participant comments from the qualitative data indicated problems with the rubric: (a) areas should be listed, (b) no direction if an inappropriate or inaccurate justification listed, (c) many samples received lower scores because all four areas were not justified, and (d) wording in cell one and two need attention. The rubric could instead state, "Justification is missing for one goal OR three or more *appropriate* justifications of the required areas in the prompt." The "appropriate" language could also be included in the Beginning level description with "Justification is missing for two goals OR less than three *appropriate* justifications of the required areas in the prompt."

Learning Goal and Pre/Post Assessment: Indicator 6 (LG 6)

A third area in the Learning Goal and Pre/Post Assessment section with a high degree of variability of a 1.05 standard deviation was the sixth indicator. The indicator states: "All assessment items are clearly and appropriately aligned to specific Learning Goals, correct level of revised Bloom's, and content standard" (School of Teacher Education, 2011, p. 12). See Table 5 for the excerpt from the TWS scoring rubric of this indicator.

Table 5
TWS Scoring Rubric Indicator Learning Goal 6 (LG6)

Indicator	Beginning	Developing	Proficient	Exemplary
LGA 6 Pre/Post Assessment Blueprint: Learning Goals	All assessment items are not aligned to specific Learning Goals, correct level of Bloom's, and content standard.	All assessment items are clearly and appropriately aligned to 2 of the following: specific Learning Goals, correct level of Bloom's, and content standard.	All assessment items are clearly and appropriately aligned to specific Learning Goals, correct level of Bloom's, and content standard.	Achieves the Proficient level with minimal assistance on the first attempt and demonstrates above and beyond the Proficient level.

The indicator brings up the challenge of the correct level of revised Bloom's which was an issue with Learning Goal and Pre/Post Assessment second indicator. If participant scorers misunderstood the revised Bloom's taxonomy, the ratings for this indicator might also be inaccurate. The qualitative data reveal lack of direction and information in the rubric such as "The word 'incorrect' should be used in cell one, not correct" and "Include the number of test items per goal in the rubric."

Learning Goal and Pre/Post Assessment: Indicator 9 (LG 9)

A fourth indicator in the Learning Goal and Pre/Post Assessment section with the highest standard deviation of 1.19 was indicator nine. The indicator states: "Scoring procedures are explained, assessment items or prompts are clearly written, mastery levels defined, directions and procedures are clear to students. Scoring key and/or rubrics are attached and include all required components" (School of Teacher Education, 2011, p. 12). Table 6 illustrates an excerpt from the TWS scoring rubric of this indicator.

Table 6
TWS Scoring Rubric Indicator Learning Goal 9 (LG9)

Indicator	Beginning	Developing	Proficient	Exemplary
LGA 9 Pre-post Assessment Blueprint: Scoring Criteria	Scoring procedures are not explained; assessment items or prompts are not written for student understanding; mastery levels are not defined; directions and procedures are not clear to students. Scoring key and/or rubrics are incomplete.	Scoring procedures are not well explained; assessment items or prompts are not clearly written; mastery levels are not clearly defined; directions and procedures are not clear to students. Scoring key and/or rubrics are attached but do not include all required components.	Scoring procedures are explained, assessment items or prompts are clearly written, mastery levels defined, directions and procedures are clear to students. Scoring key and/or rubrics are attached and include all required components.	Achieves the Proficient level with minimal assistance on the first attempt and demonstrates above and beyond the Proficient level.

The indicator includes five components (see Table 6) and it may be an issue for scorers to check for all five items. The Developing and Beginning level descriptors for the indicator state: "Scoring procedures are not well explained; assessment items or prompts are not clearly written; mastery levels are not clearly defined; directions and procedures are not clear to students. Scoring key and/or rubrics are attached but do not include all required components" (School of Teacher Education, 2011, p. 12). The Beginning descriptor states, "Scoring procedures are not explained; assessment items or prompts are not written for student understanding; mastery levels are not defined; directions and procedures are not clear to students. Scoring key and/or rubrics are incomplete" (School of Teacher Education, 2011, p. 12). Participants struggled to decide if teacher candidates had mistakes in two of the required components would they score in the Developing category, or would two mistakes cause them to receive a Beginning rating? Scoring participant comments supported this finding as one participant noted that the rubric should include a list and require a certain number instead of using language like "not well explained." If instead the rubric specified, "Assessment appropriately includes five of the required items" and the Beginning level descriptor could be, "Assessment appropriately includes four or fewer of the required items." By quantifying how many are expected, it is clearer to the scorer of what is required.

Design for Instruction: Indicator 4 (DFI 4)

The only indicator within the Design for Instruction section with a high standard deviation (.97) was the fourth indicator. The Proficient indicator states: "Thorough and clear description of two instructional strategies from different Learning Goals that includes:

Identification of appropriate content related strategies to meet Learning Goals and revised Bloom's levels; Instructional strategies meet student needs through appropriate adaptations and differentiated instruction based on pre-assessment data. Real world connections; Discussion of materials/technology" (School of Teacher Education, 2011, p. 14). Table 7 illustrates the TWS scoring rubric of this indicator.

Table 7
TWS Scoring Rubric Indicator Design for Instruction (DI 4)

Indicator	Beginning	Developing	Proficient	Exemplary
DI 4 Instructional Strategies	Provides a limited description of two instructional strategies from different Learning Goals for 2 of the following criteria in unit overview: Identification of appropriate content related strategies to meet Learning Goals and revised Bloom's levels; Instructional strategies meet student needs through appropriate adaptations and differentiated instruction based on pre-assessment data. Real world connections; Discussion of materials/technology.	Provides an adequate description of two instructional strategies from different Learning Goals for 3 of the following criteria in unit overview: Identification of appropriate content related strategies to meet Learning Goals and revised Bloom's levels; Instructional strategies meet student needs through appropriate adaptations and differentiated instruction based on pre-assessment data. Real world connections; Discussion of materials/technology.	Thorough and clear description of two instructional strategies from different Learning Goals that includes: Identification of appropriate content related strategies to meet Learning Goals and revised Bloom's levels; Instructional strategies meet student needs through appropriate adaptations and differentiated instruction based on pre-assessment data. Real world connections; Discussion of materials/technology.	Achieves the Proficient level with minimal assistance on the first attempt and demonstrates above and beyond the Proficient level.

In order to meet the Proficient rating teacher candidates must complete several parts: (a) identification of strategies to meet the Learning Goals and revised Bloom's levels, (b) appropriate adaptations and differentiation, (c) real-world connections, and (d) discussion of materials and technology. Scoring participant comments had much to say about this indicator and rubric. For example, "This seems to be a weak area for students as they don't connect to pre-assessment data. I think the prompt does not include this specific connection." One scorer

simply summed it up with this comment, "Too much information is expected." Additionally, with this section assessing alignment to revised Bloom's taxonomy, similar problems with the second and ninth indicator in the Learning Goal and Pre/Post Assessment would also be issues and may well indicate a misunderstanding of the applications of the taxonomy. A scoring participant comment supported this finding: "If Bloom's is wrong in the Learning Goal section, then Bloom's is off here, too."

Reflection of Teaching Practices: Indicator 3 (ROT3)

The third indicator in the Reflection of Teaching Practices section had a high standard deviation of .98. The Proficient description for the indicator states: "Appropriate, logical, detailed discussion of 2 of teacher's needs for improvement as related to self-evaluation of Kentucky Teaching Standards. Clearly describe 2 to 3 priorities for your own professional development based on specific data from self-assessment and student performance. Include a specific plan for growth" (School of Teacher Education, 2011, p. 24). Table 8 depicts the TWS scoring rubric of this indicator.

Table 8
Reflection of Teaching Practices (ROT 3)

Indicator	Beginning	Developing	Proficient	Exemplary
R3 Identify areas of Professional Development	Discussion of teacher's needs for improvement is not related to self-evaluation of KTS Or only one improvement is discussed. Description of one or more priorities for your own professional development is vague and not clearly based on specific data from self-assessment and student performance. Include a specific plan for growth.	Discussion of one or more of teacher's needs for improvement as related to self-evaluation of KTS may not be clear, logical, or appropriate. Description of one or more priorities for your own professional development is not clearly based on specific data from self-assessment and student performance. Include a specific plan for growth.	Appropriate, logical, detailed discussion of 2 of teacher's needs for improvement as related to self-evaluation of KTS. Clearly describes 2 to 3 priorities for your own professional development based on specific data from self-assessment and student performance. Include a specific plan for growth.	Achieves the Proficient level with minimal assistance on the first attempt and demonstrates above and beyond the Proficient level.

One potential problem with the indicator is the explicit instructions in the prompt are not included in the rubric. The prompt states, "Based on the data from the self-assessment and

student performance on this unit, identify two areas on which you need improvement. Discuss at least two types of professional development for each identified area of need" (School of Teacher Education, 2011, p. 23). Thus, the prompt calls for two types of professional development for each area whereas the rubric does not indicate the specification for two types of professional development for each identified area to improve. Some scorers might score the work as proficient if only one professional development experience is included, due to the lack of specificity in the rubric. In addition, participant scorers noted that expectations for teacher candidate responses should include more than a statement of "I will meet with a teacher." One scoring participant stated, "I interpret specific to mean that the student has done some searching to find some resources, websites, strategies, or self-help resources, but something besides just, 'I will ask my colleagues or go to a PD on this topic' and yet with the wording on this indicator, what they typically list must be counted as correct." The scorers felt that the TWS prompt and rubric need to clarify what type of professional development would be considered acceptable to alleviate confusion on this point.

Implication for Practice and Future Research

Some clear areas for revising the TWS emerged from the study, the first of which was to move from personal meaning of educational terms and constructs to shared understanding. Two indicators surfaced as a direct misunderstanding for how to apply revised Bloom's Taxonomy in planning and implementing instruction. The university faculty, teacher candidates, and research study participants need a common understanding of the application of revised Bloom's Taxonomy (Anderson et al., 2001). Some university faculty have participated in training focusing on the cognitive process identified for each of the revised Bloom's Taxonomy levels, but not all faculty. With new understanding, these faculty and scorers found that instruction and assessment often considered higher-level thinking processes are actually lower-level understandings. For example, in the revised Taxonomy, "comparing" is listed as a cognitive process in the Understanding level and not considered on the Analyzing level (Stobaugh, 2013). Some scorers incorrectly evaluated Learning Goals involving comparisons on the Analyzing level. The study unveiled internal problems in the TWS scoring process due to a lack of understanding of the hierarchy of the revised Bloom's levels and how those verbs are to be correctly used in the teaching and learning process. Therefore, a valuable initiative for all faculty and scorers would be to participate in trainings and learning opportunities focused on the revised Bloom's Taxonomy "Cognitive Processes" and learning levels.

In several areas the rubric needs to be adjusted with quantifiable words to indicate level of performance. Words like "somewhat" need to be removed and replaced with more quantifiable terms with the number of components that are required. For example, Contextual Factor 1, the School Information indicator, is an example of the use of quantifiable language: "Characteristics of school described clearly at a substantive, accurate, and unbiased level in all of the five required areas. School information provided includes the five required areas and at least one additional area" (School of Teacher Education, 2011, p. 21). Through indicating the exact number of expected skills to be demonstrated at each level on the indicator, teacher candidate responses can be more reliably scored. Finally, one indicator included several categories to be assessed in the current TWS document making it difficult to determine a teacher candidate's strengths and weaknesses on key skills. Course instructors have often indicated that differentiation has been a relatively weak area for teacher candidates. Participants in the study stated a specific desire to assess differentiation within the Design for Instruction section as its

own indicator providing a clearer picture as to whether teacher candidates can appropriately meet the diverse need of their students when planning for instruction.

Future TWS reliability research will select fewer samples that are scored by the four different groups to more closely examine the inter-rater reliability. This research indicates the need to examine differences among scoring practices of various raters: university faculty teaching the Student Teacher Seminar, professors in the education department, professors in other departments including content areas associated with TWS, and P-12 practitioners. Various groups may interpret the rubric in different ways.

To maintain quality of scoring by professors of the Student Teacher Seminar class, a regular training is needed followed by a quality control scoring session. This ensures all instructors have common expectations when scoring and rubric expectations are clearly understood. A new research study could examine the impact of the training on TWS scoring reliability.

Other professional education units may want to simulate this process to improve their culminating performance for pre-service teachers. As universities seek to maintain a low cost and research-based culminating performance for teacher candidates, other professional education units may want to assess the reliability of their performance assessment for teacher candidates.

Conclusion

The overall interesting piece of this research is the continuing issue of face validity versus construct validity. In line with face validity, faculty found that while TWS scorers agreed on the holistic score, there was a large variance on indicator scores. This led to a revision in the TWS prompts and rubrics as well as a desire to continue to improve the rubrics and assess the reliability of the document particularly by examining the indicators. What may prove to be the most important finding of the study, however, is the need to examine the differences among scoring practices of raters because scoring varies among people. Yet, even deeper is the concern with construct validity where faculty make common errors that include misinterpretation of scoring rubrics, prompts, the teaching and learning process, and even concepts such as revised Bloom's Taxonomy due to lack of shared meaning of educational terms and constructs. This finding could be generalized to other universities as all education programs utilize scoring prompts and rubrics to measure teacher candidate performance and most all use revised Bloom's Taxonomy in the teaching and learning process.

As institutions drive toward ensuring performance assessments are reliable, faculty must develop consensus around key educational concepts. City, Elmore, Fiarman, and Teitel (2009) term this as a "culture-building process." To build consistency in the instruction and assessment across faculty, critical conversations need to occur to build understanding and agreement around these concepts. Although this institution had begun the process of having conversations and making revisions to the Teacher Work Sample, it is clear that the dialogue needs to continue and keep striving forward in the difficult task of shared-meaning building.

Just as teacher candidates are required to "…collect and analyze data related to their work, reflect on their practice, and use research and technology to support and improve student learning" (NCATE, 2008, p. 19), university faculty must do the same to prepare teacher candidates for 21st century classrooms. By looking at the TWS data beyond the face reliability

perspective, institutions can pave a path for what they can discover with construct validity research that can potentially lead to meaningful and deeper change.

References

Anderson, L., Krathwohl, D., Airasian, P., Cruikshank, K., Mayer, R., Pintrich, P., Raths, J., & Wittrock, M. (2001). *A taxonomy for learning, teaching, and assessing*. New York: Addison Wesley, Longman.

City, E. A., Elmore, R. F., Fiarman, S. & Teitel, L. (2009). *Instructional rounds in Education.* Harvard, MA*:* Harvard Education Press.

Girod, G. (2002). *Connecting teaching and learning: A handbook for teacher educators on teacher work sample methodology.* Washington, DC: AACTE Publications.

Kentucky Department of Education (2008). *Kentucky teaching standards.* Retrieved from http://www.kyepsb.net/teacherprep/standards.asp.

Kohler, F., Henning, J., & Usma-Wilches. (2008). Preparing preservice teachers to make instructional decisions: An examination of data from the teacher work sample. *Teaching and Teacher Education*, *24*(8), pp. 2108-2117.

McConney, A., Schalock, M., & Schalock, H. (1998). Focusing improvement and quality assurance: Work Samples as authentic performance measures of prospective teachers' effectiveness. *Journal of Personnel Evaluation in Education*, *11*, pp. 343-363.

National Council for Accreditation of Teacher Education. (2008). *Professional standards for the Accreditation of Teacher Preparation Institutions*. Washington, DC: Author.

Nitko, A. J. & Brookhart, S. M. (2010). *Educational assessment of students* (6th ed.). Upper Saddle River, NJ: Pearson Merrill/Prentice Hall.

Popham, W. J. (2009). All about assessment: Unraveling reliability. *Educational Leadership, 66*(5), pp. 77-78.

The Renaissance Partnership for Improving Teacher Quality (2001). *Teacher work sample:*

Performance prompt, teaching process standards, scoring rubrics. Retrieved from http://fp.uni.due/itz/ProjectActivities/index.htm

Schalock, H. & Myton, D. (1988). A new paradigm for teacher licensure: Oregon's demand for evidence of success in fostering learning. *Journal of Teacher Education, 39*(6), pp. 8-16.

School of Teacher Education. (2011). *Teacher Work Sample*. Bowling Green, KY: Western Kentucky University. Available: http://www.wku.edu/teacherservices/ student_teaching/ documents/teacher_work_sample.

Stobaugh, R. (2013). *Assessing critical thinking in elementary schools. Meeting the common core*. Larchmont, NY: Eye on Education.

Stobaugh, R, Tassell, J., & Norman, A. (2012). Improving preservice teacher preparation through the Teacher Work Sample: Exploring assessment and analysis of student learning. *Action in Teacher Education*, *32*(1), pp. 39-53.

Tassell, J., Stobaugh, R., & McDonald, M. (2013). Math and science teacher candidates' use of technology to facilitate teaching and learning during student teaching. *Educational Renaissance*, 2(1), pp. 17-28.