DOCUMENT RESUME

ED 463 285 SP 040 623

AUTHOR Denner, Peter R.; Salzman, Stephanie A.; Harris, Larry B.
TITLE Teacher Work Sample Assessment: An Accountability Method That Moves beyond Teacher Testing to the Impact of Teacher Performance on Student Learning.
PUB DATE 2002-02-00
NOTE 48p.; Paper presented at the Annual Meeting of the American Association of Colleges for Teacher Education (54th, New York, NY, February 23-26, 2002). Appended material is not available from ERIC.
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Standards; Accountability; Early Childhood Education; Elementary Secondary Education; Evaluation Methods; Higher Education; *Performance Based Assessment; Preservice Teacher Education; Special Education; Student Evaluation; Student Teacher Evaluation; *Teacher Certification; Teacher Competencies; *Work Sample Tests

ABSTRACT

This paper shows how a mid-sized teacher education institution proactively developed a performance assessment method, Teacher Work Samples (TWSs), addressing the school's efforts to use TWSs to obtain evidence of the impact of teacher performance on student learning. The paper examines challenges faced in developing and implementing TWS assessments and how the school uses TWSs to hold graduates accountable for program and state standards. It also presents evidence supporting the validity and reliability of using TWSs for high-stakes assessment and program accountability. Evidence comes from over 400 work samples collected during the 2000-02 academic year. Candidates completed two TWSs in conjunction with two internships taken during the teacher education program. Overall, TWS assessments met the elements of Crocker's (1997) content representativeness (frequency, importance or criticality, and realism). The TWS measured state standards targeted in the work sample. Ratings found using the analytic scoring rubric were sufficient for making judgments regarding candidates' TWS performance. TWS performance remained constant from students' first to second internship experiences. Determination of the quality of assessment evidence was problematic. The study found that evidence of the impacts of candidate performance on student learning must be embedded within the context of the quality of the assessment evidence. Teacher Work Sample Index of Student Learning Assessment (published by the College of Education, Idaho State University) is appended. (Contains 15 tables and 19 references.) (SM)

Teacher work sample assessment: An accountability method that moves beyond teacher testing to the impact of teacher performance on student learning

Peter R. Denner

Stephanie A. Salzman

Larry B. Harris

Idaho State University

A symposium presented at the annual meeting of the American Association of Colleges for Teacher Education, New York, New York, February 23-26, 2002. Send Correspondence to Peter Denner, Box 8059, College of Education, Idaho State University, Pocatello, ID 83209 (Phone: 208 282-4230) or dennpete@isu.edu.

Teacher work sample assessment: An accountability method that moves beyond teacher testing to the impact of teacher performance on student learning

Schools, departments, and colleges of education are increasingly being held accountable for demonstrating the quality of program candidates and for documenting the impacts of their candidates on PK-12 student learning. Prompted by new federal Title II legislation requiring states to rank their teacher preparation programs, states are developing criteria to ensure that individuals entering the profession possess subject matter knowledge and the skills to help students master content presented in state standards for PK-12 students. Furthermore, the National Council for the Accreditation of Teacher Education (NCATE, 2000) has developed performance-based evaluation standards that require teacher education institutions to provide evidence of candidate knowledge, skills, and dispositions and the impact of their candidates and graduates on student learning.

In response to these federal and state mandates for accountability and new performance-based accreditation standards, many teacher education institutions have instituted assessment systems that include performance assessments such as teacher work samples, teaching performance evaluations, and portfolios (see Salzman, Denner, & Harris, 2002, this conference). These assessments are used by schools, departments, and colleges of education to make high-stakes decisions including the qualification of candidates for program admission, retention, and completion and recommendation for state teaching certification or licensure.

Institutions using performance assessments for high-stakes decisions are also faced with the challenge of showing the evidence derived from these assessments is valid and credible. As noted by Popham (1997), assessments used for high-stakes decisions such as program admission and certification or licensure must be accompanied by rigorous studies of the credibility of evidence including the validity of the assessment and the reliability of scoring decisions.

In this symposium, we show how a medium-sized teacher education institution has been proactive in the face of heightened accountability demands through the development and implementation of a performance assessment method, known as *Teacher Work Samples* (TWS).

We also address our efforts to use teacher work samples to obtain evidence of the impact of teaching performance on PK-12 student learning. Adapted from the *Teacher Work Sample Methodology* (TWSM) of Western Oregon University (Schalock, 1998; Schalock, Cowart, & Staebler, 1993), our teacher work sample performance assessment requires our teacher education candidates to document their ability to plan, deliver, and assess a standards-based instructional sequence, to profile student learning that occurred during the instructional sequence, and then to reflect on student learning in order to improve teaching practice. Important aspects of the teacher work sample are the requirements for our candidates to show how they have adapted their instruction to meet the needs of all students and to demonstrate the impacts of their teaching on student learning.

To ensure that our teacher work sample assessment responds to the mandates for program accountability and to address the technical issues of validity and scoring reliability, we greatly revised Western Oregon's TWSM. These revisions included development of guidelines for the completion of teacher work samples and a set of scoring rubrics explicitly aligned with our program standards and indicators. We further developed processes for identifying benchmarked performances, training raters, and gathering validity and reliability data. Finally, we also developed an instrument for gathering evidence of the quality of the assessment evidence used in the teacher work sample and for collecting data relative to impacts on PK-12 student learning.

The first part of the symposium addresses the primary challenges faced in developing and implementing teacher work sample assessments and the ways we use teacher work samples to hold graduates accountable for program and state standards (both based on the INTASC standards). Because this information has been presented elsewhere (Salzman, Denner, Bangert & Harris, 2001), it has not been included here. The second part of the symposium, and the bulk of this paper, presents our current evidence supporting the validity and reliability of the use of teacher work samples for the purpose of high-stakes assessments and program accountability. This evidence is drawn from more than 400 work samples collected during the 2000-2001 academic year. The final

part of the symposium and the remainder of this paper focuses on whether or not our teacher work sample assessments provide credible evidence for the impact of teacher candidate performance on student learning in terms of the quality of the assessment evidence, the percent of students who achieved the learning targets, and the number of students who showed improvement from the pre- to post-assessments.

Methods

Teacher Work Sample Assessment

Participants and Establishment of TWS Sets

Candidates complete two teacher work samples in conjunction with two internships taken during the teacher education program. The first work sample is completed during a half-time internship (junior-level) and the second is completed during a full-time student-teaching internship (senior-level). A total of $N = 411$ work samples were collected during the Fall Semester 2000 and Spring Semester 2001. Of these, $n = 150$ TWS were selected at random for benchmarking in June, 2001. Of the 150 TWS, 84 (56%) were from the initial internship candidates, and 66 (44%) were from the candidates completing their student-teaching internship. The sample of 150 TWS included all Idaho subject-area teaching endorsements and all grade levels from K to 12.

The sample of 150 TWS was subjected to a benchmarking process. Groups of qualified raters first categorized the 150 TWS along a four category developmental continuum from beginning to exemplary performance by applying a standards-based holistic scoring rubric (Salzman, Denner, Bangert, Harris, 2001). This resulted in 26 (17.3%) of the 150 TWS being classified as *Beginning*, 54 (36.0%) as *Developing*, 52 (34.7%) as *Proficient* and 18 (12.0%) as *Exemplary*.

Within these categories, groups of raters then selected TWS that were proto-typical examples of each of the categories. This resulted in the identification of a *benchmarked set* of 20 TWS (TWS Set A), consisting of 4 TWS at the *Beginning* level, 6 TWS at the *Developing* level, 6 TWS at the *Proficient* level, and 4 TWS at the *Exemplary* level. Nine of the Set A work samples

were produced by the initial internship candidates, and 11 by the student teaching candidates. The Set A TWS consisted of 10 elementary education work samples (grades K to 6) and 10 secondary education work samples (grades 7 to 12). The subject areas covered included 8 science, 6 social studies, 2 English or language arts, 2 business or computer science, 1 mathematics, and 1 consumer science.

An additional set of 20 TWS (TWS Set B) was established by selecting TWS at random from each of the 4 developmental categories. This randomly representative set of 20 TWS (TWS Set B) also consisted of 4 *Beginning*, 6 *Developing*, 6 *Proficient*, and 4 *Exemplary* work samples. In the Set B TWS, there were 10 TWS produced by the initial internship candidates and 10 produced by the candidates who were completing their student teaching internship. The Set B TWS contained 10 elementary education work samples (grades K to 6) and 10 secondary education (grades 7 to 12) work samples. The subject areas covered in the Set B TWS included 7 science, 4 social studies, 4 mathematics, 2 business or computer science, 1 English or language arts, 1 health, and 1 physical education.

Finally, from the entire group of 411 teacher work samples, we identified 40 candidates for whom we had collected both initial internship and student teaching internship work samples. This was possible during a single academic year for some of our candidates, who took their initial internship in the fall and completed their student teaching internship in the spring. Because there was a selection bias to these candidates due to the options they exercised in choosing this path through our program, the results from these sets of teacher work samples may not generalize to our entire teacher education program. However, these were the only teacher candidates from whom we had received as yet both work samples. From the 40 candidates, the TWS of 20 candidates were randomly selected and two representative sets (TWS Set C and TWS Set D) were created through the random assignment of 10 candidates to each set. Both sets (Set C and Set D) contained two TWS for each of 10 teacher education candidates, one from each occasion of development, for a total of 20 TWS in each set. The Set C TWS had 11 elementary education and 9 secondary

education work samples. The subject areas were 2 English/language arts, 8 science, 3 mathematics, 4 social studies, and 3 health. The Set D TWS had 14 elementary education and 6 secondary education work samples. The subject areas included 4 English/language arts, 3 science, 5 mathematics, 4 social studies, 2 foreign language, and 2 business.

Production and Collection of Work Samples

The standards, guidelines, and scoring rubrics employed in this study were the same as those presented by Salzman, Denner, Bangert & Harris (2001). The directions took the form of a set of *Teacher Work Sample Guidelines* (Salzman, Denner, Bangert & Harris, 2001) designed to take each candidate step-by-step through the development of the work sample tasks. The tasks required the teacher education candidates to develop a written product that included the following components: (1) a description and analysis of the learning-teaching context, (2) achievement targets for the instructional sequence, (3) an assessment plan, (4) plans for an instructional sequence comprised of at least six related learning activities aligned to the achievement targets to be taught over a four-week time period, (5) analysis of student learning, and (6) evaluation and reflection on the success of the instructional sequence with regard to student learning and future practice.

Our teacher education candidates all complete two teacher work samples during their teacher education program. For this study, some of the TWS were completed as a requirement for a junior-level course that includes a half-time internship in a PK-12 classroom. Other TWS were completed by different students during their student teaching internship. Some of the work samples included in this study were developed by candidates on both occasions During the junior-level course, the teacher candidates were assisted in the development of their first teacher work sample. The work samples completed during student teaching were completed independently by the candidates with minimum assistance from the cooperating teachers with whom they were placed. All of the work samples for this study were collected during the academic year 2000-2001.

Teacher Work Sample Scoring Rubrics

Based on the targeted standards and indicators for the TWS assessment, an analytic scoring rubric was developed (see Salzman, Denner, Bangert & Harris, 2001) for rating TWS performances on each of the six targeted standards. The analytic scoring rubric lists the targeted standards with a description of the indicators for each standard that are the criteria for judging performances relative to the standard. For example, the Reflection standard (e.g., *The teacher reflects on his or her instruction and student learning in order to improve his or her teaching practice.*) includes the following indicators: (1) draws conclusions about the extent to which the achievement targets were met and cites evidence to support those conclusions; (2) discusses questions and issues the instructional sequence raised about teaching and students; and (3) reflects on aspects of the instructional sequence that were especially successful or effective and on how the instructional sequence might be taught differently or more effectively. Each TWS is rated for all six standards on a 3-point scale: 0 = *Standard Not Met;* 1 = *Standard Partially Met*; and 2 = *Standard Met.* Summation of the ratings across the six standards yields a total analytic score.

We also used a holistic scoring rubric for making judgments regarding the total performance of our teacher education candidates on the teacher work sample assessment. The holistic rubric categorized the total TWS performance on a developmental continuum using a 4-point scale: 1 = *Beginning*; 2 = *Developing*; 3 = *Proficient*; and 4 = *Exemplary* (see Salzman, Denner, Bangert & Harris, 2001). On the holistic rubric, each category of performance was described in accordance with the degree to which performances at that level meet the standards and indicators. Thus, the holistic score marks an overall judgment made by a qualified rater, depicting the degree to which the teacher work sample provided evidence of meeting all six of the targeted standards.

Teacher Work Sample Raters

Our panel of 20 TWS raters consisted of 8 public school teachers, including 3 National Board Certified teachers, and 12 teacher education faculty members. There were 3 male raters and

17 female raters. Five of the teachers taught in elementary schools and three taught in secondary schools. The teachers averaged 13 years of teaching experience. The teacher education faculty members average 9.6 years (ranging from 0 to 32 years) of public school teaching experience and 8.3 years (ranging from 1 to 22 years) of college teaching experience.

Procedures for Teacher Work Sample Scoring

The benchmarked TWS used for this study were originally identified in June, 2001. The one-day session started with two hours of training. The training covered the purpose of the benchmarking process, the TWS standards, the guidelines the teacher candidates' used to develop their work samples, and an extensive review of the scoring rubrics. We also conducted anti-bias training for uncovering potential scoring bias due to personal preferences regarding an ideal teacher work sample. As part of this training, raters were first directed to list characteristics of excellent teacher work samples. The raters then compared the characteristics they wrote on their personal lists to the standards targeted in the work sample. If a listed characteristic did not appear in the standards, then the raters were instructed to record it on their *Hit List of Personal Biases*. These lists were used as references during the scoring process to remind the raters to focus on only the standards and indicators when rating the teacher work samples. As a final part of the training preparation for the benchmarking activities, the raters were given additional directions regarding respect for the confidentiality of the performances, the security of the teacher work samples used in the study (they were not to leave the building), the importance of avoiding halo and pitchfork effects in scoring, and the importance of searching for evidence throughout the entire work sample.

The 16 raters were then divided into 5 groups. The task of each group was to reach consensus on the holistic score category of their assigned TWS and place each work sample in one of four piles representing the four levels of the scoring rubric. Each group performed a *quick read* of 30 (20%) of the 150 work samples. This process, which took about 2 hours, resulted in all 150 work samples being distributed to one of the four holistic score categories. In the afternoon, the same 16 raters were assigned to 4 different groups. Each group was assigned the task of choosing

four to six examples of TWS performances at one level of the four holistic score categories. This process also took about two hours and resulted in the selection of 20 benchmark TWS. This set of 20 TWS consisted of 4 *Beginning*, 6 *Developing*, 6 *Proficient*, and 4 *Exemplary* work samples (This set became TWS Set A). Several additional benchmark examples were identified for use in the training of future raters. After the raters had completed the task of selecting benchmark examples, an additional 20 TWS were selected at random from the remaining TWS in each of the four holistic categories (This set became TWS Set B). This second non-benchmarked but representative set of TWS also had 4 *Beginning*, 6 *Developing*, 6 *Proficient*, and 4 *Exemplary* work samples.

During the scoring session in October, 2001, the same two hour training was repeated once more for all of the raters. Some of the raters who participated in the October scoring session had also participated in the earlier benchmarking session in June. Other raters were new to the process of scoring teacher work samples. The training consisted of a review of the TWS standards, guidelines, and scoring rubrics. We also conducted a new round of anti-bias training. The initial training again took about 2 hours. The raters next practiced scoring two teacher work samples. The first scoring was a guided group scoring of a benchmarked TWS. After discussion, this was followed by an individual scoring of a second benchmarked TWS. The total scoring practice with discussion took about an hour and a half. In the afternoon, a second individual practice scoring was completed using a third benchmarked TWS. After this, we had the raters complete the validity questionnaire and a demographic questionnaire. Together these activities took about 2 hours.

Around the middle of the afternoon, the raters were randomly assigned to six groups for independent scoring of the TWS sets. Two of the six groups had 4 raters and four groups had 3 raters. The two 4 rater groups were randomly assigned to analytically score either TWS Set A or TWS Set B. The four groups of 3 raters were randomly assigned to score either TWS Set C or TWS Set D and to scoring method (either analytic or holistic). Each rater was given a box of 20 TWS containing their assigned set of work samples along with a sufficient number of scoring

rubrics for their assigned method of scoring. The 4 groups of raters assigned to score TWS Set C and TWS Set D were not told their boxes contained teacher work samples developed by the same teacher candidates across two occasions. Information about the candidates and the occasion of development had been removed from all of these work samples.

Before releasing the raters to take their boxes of work samples home to score, the raters were told the importance of scoring the TWS independently without any discussion with the other raters. They were also advised to review their bias hit lists each time they began to score. We also asked them to record their start and stop times for each scoring. Scoring time was investigated as part of this study and will be presented in the results section for both the analytic and holistic scoring rubrics. Finally, all of the raters were asked to return their completed boxes as soon as possible, but no later than December 1, 2001.

Teacher Work Sample Validity Questionnaire

We applied Linda Crocker's (1997) methodology for performing content judgments of performance assessment exercises and scoring rubrics by developing a teacher work sample validity questionnaire. The questionnaire first asked our panel of raters to evaluate the degree of alignment between the TWS standards, guidelines, and scoring rubrics. We next asked the panel of raters to evaluate the degree to which Crocker's criteria for content representativeness were met. The criteria included the *frequency* of the teaching behaviors in actual job performance, *the importance or criticality* of those behaviors, of the behavior, the authenticity (or *realism)* of the tasks to actual classroom practice and the degree to which the tasks represented the targeted standards. The raters were also asked to assess the degree to which the elements of the teacher work sample and the scoring rubric assessed each of the ten Idaho Core Teacher Standards (Idaho State Board of Education, 2000). Finally, the raters were asked to respond to four questions about the overall validity of the teacher work sample assessment, including whether they considered it appropriate to use teacher work samples as an accountability measure for demonstrating beginning teacher competence.

Design for Assessing Rater Agreement on the TWS Scoring Rubrics

We examined the issue of inter-rater agreement using concepts from Generalizability Theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991). Formulas supplied by Shavelson and Webb (1991) were used to calculate coefficients of dependability, which are similar to reliability coefficients in classical test theory. These same formulas were adjusted to provide information regarding the number of raters necessary for making high-stakes decisions about absolute teaching performance level. Two designs were used in this study. The first design examined only a rater facet to access the effect of rater on the generalizability of scores based on our analytic scoring method for benchmarked (TWS Set A) and non-benchmarked (TWS Set B) sets of work samples. For the second design, in addition to a *rater* facet, we examined the facet of *occasion.* Because we had collected two work samples from some of our teacher education candidates (one during their initial internship and one during their student teaching internship), we were able to examine the effect of development *occasion* as a facet. Both of these designs and the methods used for analyzing them are explained in greater detail in the results section.

Impact on Student Learning

Panel of Expert Raters

To examine whether or not our teacher work sample assessment provided credible evidence for the impact of teacher performance on student learning, we had a panel of raters apply multiple criteria for assessing the quality of evidence contained in the work samples. Our panel of expert raters was composed of three male education faculty members. These faculty members did not have prior experience with helping our teacher education candidates to develop teacher work samples. All of them had extensive educational background in tests and measurements. All of them had background in teaching graduate level courses and workshops on assessment and measurement, and all had experience in the development and use of scoring rubrics. They averaged $M = 11$ years of university teaching experience.

Index of Student Learning

To assess the quality of evidence for student learning contained in the work samples, including such factors as the quality of the assessment items, the percent of students who achieved the learning targets, and the number of students who showed improvement from the pre- to post-assessments, we developed an *Index of Student Learning* measure with three parts. The first part was composed of 13 items covering the *Quality of Sources of Evidence* for the assessments employed in the work samples. Each item focused on a different assessment criterion, such as whether the achievement targets were appropriate for the grade level or whether the assessments directly measured the achievement targets (see Appendix A). Each of the 13 items was rated as: 0 *Does Not Meet Criterion*, 1 *Partially Meets Criterion*, or 2 *Meets Criterion*. A total score was obtained by summing the ratings across the 13 items.

Part two of our Index of Student Learning asked the raters to determine the percent of students who met the stated achievement targets in each TWS by applying the criterion for success level set in the TWS itself. To determine an answer to this question the raters either found this information in the TWS or tried to use the data available in the TWS to calculate a percent. If they could not determine a percent, then they responded that the data were not available. Part three of our Index of Student Learning asked the raters to determine the percent of students who showed improvement relative to the learning targets in the TWS. Once more, if the raters could not locate or determine a percent, then they responded the data were not available.

Procedures for Examining Impact on Student Learning

To investigate the validity of teacher work samples as an assessment of the impact of teacher performance on student learning, we had three faculty members with expertise in measurement examine our benchmarked set of teacher work samples (TWS Set A). The three raters for this portion of our study first participated in an extensive full-day training session. The training included review of the teacher work sample guidelines and targeted standards. The raters also spent several hours reviewing and discussing the Quality of Sources of Evidence items on the Student

Learning Index to develop shared definitions of terms and common understanding of the sources of evidence for each item and where in the teacher work sample the evidence might be found. During this process of reviewing and discussing each item, any personal biases regarding teaching and assessment that were uncovered were isolated and recorded (in a manner similar to the anti-bias training described previously for the TWS raters). These lists of personal biases were used by the raters during actual scoring to remind them to focus on the criteria stated in the student learning index items only.

As the next step in the training for scoring, the raters reviewed the possible sources of information for answering the questions about the percent of students achieving the achievement targets and the percent of students demonstrating improvement. Methods for computing the percent of students making improvement toward or reaching the achievement targets for the instructional sequence were also discussed and where the necessary information for computing percentages might be found in the work samples. As part of this process, raters reached consensus regarding how to rate teacher work samples that contained no data relative to student learning. Finally, the raters scored a teacher work sample and discussed their ratings in order to calibrate scoring decisions.

Upon completion of this training, the three raters were each given a set of 20 teacher work samples (Set A) and copies of the Student Learning Index to rate them. The raters were instructed to score the teacher work samples with the Student Learning Index using the definitions of terms and scoring directions developed during the training. The raters scored the teacher work samples individually over a four-week period. The expert raters took an average total of $M = 14.1$ minutes ($SD = 11.25$ minutes) to score each work sample using all three parts of the Index of Student Learning.

Results

Validity

Alignment

To support alignment, we had the panel of raters ($n = 20$) evaluate the relationship between our TWS Guidelines, the targeted TWS standards, and the analytic scoring rubric. Table 1 presents the judgments made by our panel of 20 raters. The data were missing from one rater on two of the alignment statements. All responses indicated the raters thought there was a moderate to high degree of alignment among the guidelines, the standards, and the scoring rubric. The highest rated alignment was between the analytic scoring rubric and the targeted standards, with 89.5% of the raters indicating a high degree of alignment. The lowest rated alignment was between the task elements presented in the TWS guidelines and the targeted TWS standards. Yet, even here, the alignment was considered high, with 80% of the raters indicating a high alignment. Together, these data support the criteria of alignment among standards, rubrics, and tasks necessary for valid performance assessment.

Content Representativeness

We also sought support for the validity of the judgments about teaching performance made on the basis of the teacher work sample assessment. To assess content validity, we applied criteria suggested by Crocker (1997) for judging the *content representativeness* of performance assessments and rubrics. These criteria included the *frequency* of the targeted behaviors in actual practice, the importance or *criticality* of the targeted behaviors to real performance, the *authenticity* of the tasks to actual performance situations, and the *representativeness* of the tasks with respect to the targeted performance standards. Each of these criteria will be considered in turn.

Frequency. Our panel of 20 raters were asked to indicate how frequently they would expect a teacher to engage in the teaching behaviors targeted by the teacher work sample assessment. Table 2 presents the judgments made by our panel of 20 raters. Across the teaching behaviors, 75% to 100% of the raters indicated a high frequency of weekly or daily for all of the targeted

teaching behaviors. These results support the frequency criteria of content representativeness. The highest rated teaching behavior in terms of its frequency (95% said daily) was reflection on instruction and student learning in order to improve teaching. The lowest rated teaching behavior in terms of frequency was the use of assessment data to profile student learning, communicate information about student progress, and plan future instruction. However, 75% of the raters still considered this latter activity to be weekly (50%) or daily (25%). Thus, all of the targeted teaching behaviors were considered to have a high frequency in actual teaching practice.

Importance. Table 3 presents the number and percent of the raters indicating the importance of the teaching behaviors targeted by the TWS assessment to effective teaching. As can be seen in Table 3, all of the teaching behaviors were considered to be important or very important, with 75% or more of the raters indicating very important across all the teaching behaviors. Thus, teacher work sample assessments satisfy the importance criteria of content representativeness.

Authenticity. Our panel of raters ($n = 20$) was next asked to judge how authentic the tasks required by the TWS are to success as a classroom teacher. Table 4 presents the ratings for each of the seven major TWS tasks. The majority (60% or higher) of the raters said all but one of the tasks are authentic or very authentic. The major exception was asking teacher education candidates to "summarize student learning, including graphs or charts that profile student performance on pre-assessment and post-assessments, and disaggregate assessment data to analyze trends or differences in student learning." Only 40% of the raters thought this task was authentic or very authentic. Comments made by the raters indicated they believed that teachers do not regularly summarize student learning in this way, even if it might be a good idea for them to do so. This task, however, is critical if teacher work samples are to be used to demonstrate the impact of teaching performance on student learning. Overall, these results support the authenticity criteria for the content representativeness of our TWS assessment.

Representativeness. We also asked our panel of raters to consider the degree to which the tasks required by the teacher work samples reflect and represent the targeted standards. The ratings

for the seven major tasks required by the work samples are presented in Table 5. All of the tasks were thought to be representative or very representative of the standards by the vast majority of the raters (95% to 100%). Hence, this criteria for the content representativeness of our TWS performance assessment was also supported.

Overall validity. We additionally asked the raters to respond to four questions concerning the overall validity of teacher work samples as assessments of teaching performance. The first question was, "Overall, does the Teacher Work Sample measure knowledge and skills that are *necessary* for a beginning teacher?" The responses of the raters were $f = 6$ (30%) said they are necessary and $f = 14$ (70%) said they are absolutely necessary. The second question was, "Overall, *how critical* to the practice of a beginning teacher are the teaching skills and strategies the Teacher Work Sample requires teacher candidates to demonstrate?" To this question, $f = 9$ (45%) said critical, and $f = 11$ (55%) said absolutely critical. The third question asked, "Overall, does the Teacher Work Sample present teacher candidates with *realistic* performance situations similar to ones they might encounter in professional practice as a teacher?" This time $f = 9$ (45%) of the raters said somewhat realistic, $f = 7$ (35%) said realistic, and $f = 4$ (20%) said absolutely realistic. The raters who thought the performance situations were only somewhat realistic said that the main problem was time--"Teachers do not have enough prep time to do this in this detail." Our final question was, "Overall, how *appropriate* is it to use a Teacher Work Sample as one measure of a beginning teacher's competency?" Eight of the raters (40%) said appropriate, and $f = 11$ (55%) of the raters said it was absolutely appropriate. Only one of the raters chose the response of somewhat appropriate. This rater thought the teacher work sample assessment expected too much of teacher education candidates. All together, the raters' responses to these questions support the validity of the use of teacher work samples as a performance assessment for judging beginning teaching competence with respect to targeted program and state standards.

Alignment With State Standards

Our final content validity consideration was the degree to which the performances on the teacher work samples directly assessed any of the Idaho Core Teacher Standards (Idaho State Board of Education, 2000). The teacher work sample standards were written to reflect both program and state standards, although only some of the standards were targeted directly. Consequently, we asked our panel of experts to indicate the extent to which the tasks required for the teacher work sample measured each of the 10 Idaho Core Teacher Standards (Idaho State Board of Education, 2000) using a scale of (1) *Not At All*, (2) *Implicitly*, and (3) *Directly*. Table 6 presents the number and percent of the responses for each standard. As can be seen from the table, 65% or more of the raters said that Idaho standards 3, 4, 7, 8, and 9 were directly measured by the teacher work sample assessment. These standards were the five targeted standards for the teacher work sample assessment. Also from Table 6, it can be seen that some Idaho standards were not considered to be directly measured. Idaho standards 2, 5, 6, and 10 were judged to be implicitly measured or not at all. These standards were not targeted by the teacher work sample assessment and so should only have been seen as implicitly measured at best. The overall pattern of rater responses supports the content validity of our teacher work sample assessment as a measure of our candidates' abilities to meet the targeted Idaho Core Teaching Standards.

Generalizability

Analytic Scoring Method

To investigate the amount of variance in the total analytic scores caused by differences among raters and to determine the generalizability of our TWS scores across raters, we applied a research design from *Generalizability Theory* (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991). Generalizability Theory (Shavelson & Webb, 1991) also provides a means for determining summary coefficients reflecting the dependability of raters that is similar to classical test theory's reliability coefficient. The design was a single-facet, crossed, random-effects design. Eight raters were assigned randomly to score 2 sets of 20 TWS using the analytic scoring

rubric with 4 raters assigned to each set. Set A consisted of exemplar work samples, Set B consisted of TWS that were selected at random from the four holistic score categories. The effect for rater in this study was a random effect, because the raters were assigned at random to the two sets of TWS and they were considered to be interchangeable with other qualified raters. The study design was crossed because all 4 raters assigned to each TWS set scored all 20 work samples in their assigned set.

The above design was analyzed separately for each of the two sets of TWS using repeated measures ANOVA. The *rater facet* served as the repeated-measures factor. Table 7 presents the analysis of variance for each set. For both sets, the effect of rater was statistically significant at the .05 level of significance. Unfortunately, these results mean the scores assigned by the raters within each set of TWS differed significantly on average. In each case, post hoc mean comparisons using the Tukey method revealed one rater differed significantly from two or more of the other raters. In both instances, this rater was a teacher education faculty member with less experience scoring TWS, who scored the TWS lower on average than the other raters. This rater differed from the other teacher education faculty members and the classroom teacher. The other raters did not differ from each other. Table 8 presents the variance components used in the formulas for computing the dependability coefficient for each TWS set. The variance components are the amount of score variability attributable to each source--persons (teacher candidates), raters, and residual measurement errors.

Generalizability theory provides separate coefficients for computing the generalizability of test scores depending upon whether the measure is to be used to make decisions about the "relative standing or ranking of individuals..." or about "...the absolute level of their scores" (Shavelson & Webb, 1991, p. 84). Because we want to be able to use the analytic rubric to make decisions that reflect our candidates' abilities to meet the TWS standards (an absolute decision about performance levels) that can be generalized over raters, we used the formulas for computing an *index of dependability* for absolute decisions provided by Shavelson and Webb (1991). An index of

dependability in this case is a measure of the extent to which total scores on the analytic rubric reflect our candidates' abilities to meet the TWS standards (after potential measurement errors, such as differences due to raters, are taken into account). The same formulas also allow for estimates of dependability to be made for a different number of raters than were actually used in this study.

The 4 rater coefficient of dependability for the Set A exemplar work samples was computed to be .80 and the 4 rater coefficient of dependability for the Set B representative work samples was computed to be .73. As might be expected, the coefficient for the benchmarked TWS set was higher than for the non-benchmarked set. Fortunately, both of these coefficients are sufficient for making judgements about candidate performances. Table 9 displays dependability coefficient estimates for different numbers of raters for both teacher work sample sets. As can be seen from the table, to achieve sufficient inter-rater agreement for performance levels to be generalizable across raters, we are going to need to use panels of raters with three or more raters.

Effect of Occasion on Analytic Rubric Scores

To further investigate the generalizability of scores based on our analytic scoring method, and to examine the generalizability of the scores across scoring occasions, we developed a two faceted generalizability design with raters serving as a random facet and occasion as a fixed facet. The occasion facet was considered fixed because we only collect teacher work samples on two fixed occasions in our teacher education program, once during a junior-level internship and again during their student teaching internship. We were interested in determining whether decisions about performance levels could be generalized across raters when averaged across our two fixed measurement occasions. We were also interested in determining whether or not the measurement occasion revealed differences in performance levels. During the junior-level internship our teacher education candidates develop their work samples with the guidance and assistance of their course instructors. During the student teaching internship, our senior-level candidates are expected to following the guidelines and to produce their teacher work samples with only minimum assistance

from their cooperating teachers. Thus, we were interested in determining whether the amount of assistance made a difference to the level of performance of our teacher candidates on our analytic scoring rubric.

The two faceted design was analyzed separately using repeated measures ANOVA for each of two different sets of TWS (Set C and Set D) selected from among the candidates from whom we had collected a pair of work samples longitudinally. The Set A and Set B work samples examined previously were collected cross-sectionally. Groups of randomly assigned raters ($r = 3$) evaluated either the Set C or Set D TWS for the same 10 teacher education candidates across both occasions of TWS development. Thus, the candidate performances (persons) were crossed with both the rater and occasion facets, and the rater facet was crossed with the occasion facet.

Table 10 presents the repeated measures ANOVA for each set (Set C and Set D) for the analytic scores assigned by the three raters of each set. For TWS Set C, the effect of rater was not statistically significant at the .05 level of significance, but for TWS Set D the effect of rater was statistically significant ($p = .004$). For TWS Set D, post hoc mean comparisons using the Tukey method revealed that one rater differed from one of the other two raters. This rater was a teacher education faculty member who was scoring TWS for the first time, and who scored TWS lower than the other raters. For both sets of teacher work samples, neither the effect for occasion nor the rater by occasion interaction had a statistically significant influence on the analytic scores at the .05 level of significance. This means the analytic performances of our candidates were similar across the two occasions of measurement and also that rater differences remained constant across occasions as well. These findings indicate our candidates performed about the same when asked to complete a teacher work sample on their own (occasion 2 = student teaching) as they did on the first occasion when they received assistance from a course instructor (occasion 1 = junior-level internship with extensive training in work sample development).

Table 11 presents the variance components for Set C and Set D TWS when scored using our analytic scoring rubric. These variance components were used in the formulas for computing

dependability coefficients supplied by Shavelson and Webb (1991). The variance components represent the variance due to persons (candidate performances), raters and residual measurement errors when averaged across the two fixed measurement occasions. The three rater coefficient of dependability for TWS Set C was computed to be .86 and for TWS Set D it was computed to be .73. Again, both coefficients are sufficient for generalizing the averaged scores across occasions made by the three rater panels. They are also similar to the coefficients obtained for the cross-sectional sets previously (TWS Sets A & B).

Holistic Scoring Method

We next examined the generalizability of scores using our holistic scoring method. Applying the same two facet design, we had two additional panels of 3 raters each score the Set C and Set D teacher work samples. A group of randomly assigned raters ($r = 3$) evaluated either the Set C or Set D TWS for the same 10 teacher education candidates across the two occasions of TWS development. This allowed us to examine the generalizability of judgments made about teaching performance using our holistic scoring method across both raters and occasions of measurement. We were interested in determining whether or not the measurement occasion revealed differences in performance levels when those performances were judged holistically. As before, candidate performances (persons) were crossed with both the rater and occasion facets, and the rater facet was crossed with the occasion facet. Once again, the occasion facet was treated as a fixed effect because we only collect teacher work samples on the two fixed occasions described previously.

The above two facet design was analyzed separately for the Set C and Set D teacher work samples using repeated measures ANOVA and the holistic score ratings made by the panels of three raters. The results are presented in Table 12. For TWS Set C, the effect of rater was not statistically significant ($p = .72$), but for TWS Set D, the effect of rater was statistically significant ($p < .05$). Again, for TWS Set D, post hoc mean comparisons using the Tukey method revealed the ratings made by one rater stood out from the others. This time the different rater was a classroom teacher who was scoring TWS for the first time and who scored higher than the other raters. For

both sets of teacher work samples, neither the effect of occasion nor the rater by occasion interaction was statistically significant. This means the holistic performances of our candidates were similar across the two occasions of measurement. It also means the rater differences remained constant across occasions as well. The lack of a statistically significant effect for the occasion facet means our candidates performed about the same on the holistic rubric on the second occasion (during student teaching) as they did on the first occasion (during their junior-level internship with extensive training in TWS development).

Table 13 presents the variance components for Set C and Set D TWS when scored using our holistic scoring rubric. These variance components are the variance estimates for persons (candidate performances), raters and residual measurement errors when averaged across the two fixed measurement occasions. These variance components were used in the formulas for computing dependability coefficients supplied by Shavelson and Webb (1991). The three rater coefficient of dependability for TWS Set C was computed to be .80 but for TWS Set D it was computed to be only .39. This means the proportion of candidate holistic score differences averaged across occasions that can be generalized across raters was sufficient to generalize the holistic judgements made for the Set C TWS but not for the Set D TWS. The reason for this inconsistency appears to be the fact that the TWS Set D raters included a classroom teacher who was inexperienced in scoring work samples. This rater's scores differed from the other two raters and thus lowered the dependability coefficient for the Set D work samples. This also means rater experience mattered quite a bit to the dependability of holistic ratings. It would seem that only experienced raters should be used when making absolute decisions about candidates' levels of teaching performance using our holistic scoring rubric. This finding may also reflect the attributes of the Set D teacher work samples, because they were somewhat harder to score dependably using our analytic scoring method as well.

Scoring Time

For both the analytic scoring method and the holistic scoring method, we also examined the amount of time it took the raters to score the work samples. Table 14 presents the mean and standard deviation of the scoring times in minutes for each of the TWS sets by scoring method. The average time for scoring the benchmark work samples (Set A) using the analytic scoring rubric was 14.3 minutes. This average time is very similar to the average time of 13.5 minutes reported by Salzman, Denner, Bangert and Harris (2001) for analytically scoring benchmarked teacher work samples. As expected, it took a bit longer, $M = 24$ minutes, to score the non-benchmarked TWS (Set B) using the analytic scoring method. Similar average times were also found for Set C and Set D TWS, $M = 22.7$ and $M = 34.0$ respectively, when scored using the analytic rubric. Taken together, these results suggest a teacher work sample can be scored in about a half hour or so. This is an amount of time that is both reasonable and practical.

Surprisingly, the average times for scoring teacher work samples using the holistic scoring method varied quite a bit across the two sets of work samples (Set C and Set D). The average time of 14.7 minutes for Set D was close to the average time for scoring benchmark work samples. In contrast, the average time of 27.9 minutes for Set C was closer to the time it took to score the Set D work samples using the analytic method. These difference are hard to explain. They are most likely due to differences in the ability of the raters across the two groups to apply the holistic method of scoring to the particular set of work samples they were assigned to rate.

Fortunately, correlational analyses showed scoring time was not significantly ($\alpha = .05$) correlated with total analytic scores for any the teacher work sample sets, $r = .-.13$, $n = 77$, $p = .25$ for Set A, $r = .09$, $n = 78$, $p = .41$ for Set B, $r = .09$, $n = 39$, $p = .59$ for Set C, and $r = .17$, $n = 39$, $p = .31$ for Set D. This was true also for the correlation of scoring time with the holistic scores, $r = .21$, $n = 60$, $p = .11$ for TWS Set C, and $r = .25$, $n = 58$, $p = .06$ for TWS Set D. These results show that scores did not vary as a function of scoring time.

Evidence for Impact on Student Learning

Interrater Agreement for the Quality of Sources of Evidence Ratings

For part one of our *Index of Student Learning*, we computed a dependability coefficient for absolute decisions based on the total scores obtained across the 13 *Quality of Sources of Evidence* items made by our panel of 3 expert raters for the Set A TWS ($n = 20$). Using repeated measures ANOVA, the effect for rater was found to be statistically significant, $F(2, 34) = 34.08$, $MSE = 10.58$, $p < .001$. Table 15 presents the variance components used in the formula for computing the dependability coefficient. The three rater coefficient of dependability for total scores was calculated to be only .01. Unfortunately, this indicates near zero differentiation of candidate performances that can be generalized across raters using our quality of sources of evidence scale. As a result, no further efforts were made to relate performances on this measure to our candidates' performance on their TWS.

For all three raters, the most frequent response to the questions concerning the percent of students who met the achievement targets (part two) and the percent of students who showed improvement relative to the achievement targets (part three) was that the data were not available. For the percent of students who met the achievement targets, rater 1 said 20 out of 20 TWS lacked this information, rater 2 said it was not available for 19 out of 20 of the TWS, as did rater 3. For the percent of students who showed improvement, two of the raters said this data was not available for 19 out 20 of the TWS. The third rater said this data was not available for 12 of the 20 TWS, but also said he struggled to find the information to determine a percent for those TWS for which he said the data were available. The one thing these raters were able to agree upon was that these data were not available or was very difficult to obtain from nearly all of our candidates' work samples. This finding has important implications because it points to a need to improve our guidelines and task prompts for producing teacher work samples. It also suggests that we may need to alter our teacher preparation program to better prepare our candidates to supply this data, if our TWS are to supply credible quantitative evidence for our candidates' impact on student learning.

Discussion

The work presented in this symposium addresses the challenges faced in developing and implementing teacher work samples as a valid and credible method for documenting candidate performance relative to institutional and state standards and for providing evidence of candidate impact on PK-12 student learning. The study yielded important data regarding the validity of the teacher work sample for assessing teaching performance, alignment of the teacher work sample tasks with state standards, scoring generalizability, candidate performance across occasions, and the use of the teacher work sample as evidence of the impact of candidates on the learning of the students they teach.

Validity of the Teacher Work Sample Assessment

Data from this study indicate the teacher work sample assessment meets the elements of Crocker's (1997) *content representativeness* including frequency, importance or criticality, and realism. In terms of frequency, the expert raters determined the tasks embedded in the teacher work sample would be completed by teachers very frequently--weekly or even daily. The expert raters also judged the tasks in the work sample as critical to the job performance of teachers. Finally, the expert raters agreed the teacher work sample tasks represent actual classroom practice and strongly represent the targeted standards.

In terms of overall validity of the teacher work sample tasks as assessments of teaching performance, the expert raters agreed the tasks were necessary or absolutely necessary to the job of teaching. All of the raters also agreed the teaching skills and strategies required by the teacher work sample are critical or absolutely critical to the practice of a beginning teacher. In addition, the expert raters agreed the teacher work sample presents teacher candidates with realistic performance situations similar to ones they might encounter in professional practice as teachers. Finally, the raters agreed the teacher work sample overall is an appropriate measure of a beginning teacher's competency.

These results indicate teacher work sample assessment has great promise for providing credible evidence of the teaching performance of beginning teachers. As such, data from the teacher work sample will prove to be a powerful method for providing the evidence of candidate performance required by new accreditation standards (NCATE, 2002).

The finding that the practicing teachers on our expert rater panel judged the teacher work sample to be authentic and representative of the work of teachers has important implications for the use of the teacher work sample. Because practicing teachers view the teacher work sample as being relevant to their work, the assessment could prove to be a valuable professional development tool. The teacher work sample has the potential for providing structure for practice-based learning (Ball & Cohen, 1999; Sykes, 1999) through which professional development is embedded in teacher's everyday practice. In fact, our preliminary work (Denner, Salzman & Bangert, in press) using teacher work samples as the basis for professional development have been very positive. Through completing the teacher work sample, teachers learn new strategies and approaches for connecting their learning goals to state standards, developing and using assessments, profiling student learning, and reflecting on their practice for instructional improvement.

Alignment with State Standards

When asked to rate the extent to which the teacher work sample addressed state standards, the expert raters judged that the teacher work sample did indeed measure the state standards targeted in the work sample. Adding credibility to this judgment was the fact that the raters did accurately identify the state standards not specifically targeted in the teacher work sample. These results support the credibility of the teacher work sample as a measure for assessing candidate performance relative to state standards. Because the Idaho Core Teacher Standards are a re-statement of the INTASC standards, the conclusion could be generalized to the use of teacher work samples to assess candidate performance relative to national standards. The growing emphasis on state and national standards for teachers, characterized by Marzano and Kendall (1998, p. 4) as *"omnipresent,"* reflects an emerging emphasis on how to reform teaching and learning in schools

by improving teacher quality. As part of this reform, teacher education institutions are expected to document candidate performance relative to the knowledge, skills, and dispositions embodied in the standards (National Commission on Teaching and America's Future, 1996; NCATE, 2000).

Generalizability

The generalizability coefficients resulting from this study indicate that ratings made using the analytic scoring rubric are sufficient for making judgments regarding candidate performance on the teacher work sample. This finding is in concert with the results of Denner, Salzman and Bangert (in press). However, our data indicates that to achieve sufficient inter-rater agreement, panels of at least three raters or more are needed. It appears from this data that to achieve credible evidence about candidate performance from teacher work samples, the performances must be evaluated by multiple trained raters. Our findings also indicate that experienced raters show greater consistency with other raters than inexperienced ones. On the occasions when the effect of rater was statistically significant, it was always the least experienced rater that stood out from the others by scoring the work samples either higher or lower than the other raters. This finding has important implications for teacher education institutions using teacher work samples for making decisions regarding program retention or completion. Without doubt, if used for high-stakes decisions, the teacher work sample should be scored by multiple raters who are experienced raters that have demonstrated the ability to score work samples dependably.

An additional positive outcome of the use of multiple raters is the potential for collaborative work between university faculty and practicing educators focused on the assessment and improvement of teaching and learning. As noted by Hawley and Valli (1999), effective teacher preparation and professional development is collegial and collaborative. When university faculty and practicing educators work together, they develop a shared vision for good teaching, challenge one another's perspectives, and create mutual respect. Through this collegial collaboration, stimulated by the teacher work sample assessment and scoring process, we create a learning community focused on the learning and well-being of PK-12 students.

Performance Across Occasions

Comparisons of the teacher work samples of candidates in the pre-internship and candidates in the student teaching internship show that the performances remained constant across the two occasions of teacher work sample development. It appears from these data our candidates' performances on the teacher work sample remain stable from the first internship experience to the second experience. Should future studies support this conclusion, the teacher work sample could be used as one predictor of performance in student teaching. As such, the teacher work sample could serve as a very useful tool for qualifying candidates for the student teaching internship and for identifying students who need remediation relative to program standards prior to progressing to the final internship.

Evidence of Impact on Student Learning

Data from the Student Learning Index portion of this study indicate that the determination of the quality of assessment evidence is problematic. Our assessment experts were unable to agree on their ratings of the quality of achievement targets and assessments used in the teacher work samples. One explanation for this finding is that the evaluation of assessment evidence needs to be situated within the context of the teacher work sample. Perhaps using raters who had no experience with the teacher work sample biased the findings because the raters had insufficient context from which to judge the quality of the evidence embedded in the teacher work samples. This hypothesis is supported by the findings of a similar study (Salzman & Denner, 2002) with a national sample of teacher work samples from the Renaissance Partnership for Improving Teacher Quality in which raters did achieve reliable ratings using the Student Learning Index. In that study, expert raters consisted of measurement experts who had experience with teacher work samples. They may also have been better qualified raters. All of the raters for the Salzman & Denner (2002) study served as the assessment directors or coordinators for their respective colleges.

In terms of evidence of student learning provided in the teacher work samples, the expert raters were unable to locate reliable data showing the percent of students meeting the achievement

targets or showing improved learning relative to the targets. The expert raters did, however, find rich descriptive and interpretive data in the teacher work samples through which candidates drew accurate conclusions regarding the extent of student learning and implications for future practice. These findings indicate that the lack of data about student learning may be a function of the nature of the prompt rather than the skills and abilities of candidates. As a result of this finding, we have already revised our prompt for the teacher work sample to provide explicit directions for statistically representing PK-12 student learning. This issue has also become a focus of curriculum development work in the college to ensure candidates receive the course work and clinical experiences that will support development of the skills necessary to profile student learning.

Conclusion

Through the study presented in this symposium, we provide processes for addressing the continuing challenges faced by teacher education institutions as they strive to meet federal and state mandates for accountability and performance-based accreditation standards. This research contributes to the growing body of research (Danielson, 1996; National Board for Professional Teaching Standards, 2001) regarding the development of credible standards-based assessments of teacher performance relative to state and national standards. We also found that evidence of the impacts of candidate performance on PK-12 student learning must be embedded within the context of the quality of the assessment evidence. Moreover, if we are to prepare candidates to effectively teach in standards-based accountability systems, our teacher education programs must focus on preparing candidates to set learning goals, develop multiple assessment methods to assess student learning before, during, and after instruction, and to represent PK-12 student learning clearly and accurately.

References

Ball, D., & Cohen, D. (1999). Developing practice, developing practitioners: Toward a practice based theory of professional education. In Linda Darling-Hammond and Gary Sykes (eds.), *Teaching and learning as a profession* (pp. 3-32). San Francisco: Jossey-Bass.

Crocker, L. (1997). Assessing content representativeness of performance assessment exercises. *Applied Measurement in Education, 10*, 83-95.

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

Denner, P. R., Salzman, S. A., & Bangert, A. W. (in press). Linking teacher assessment to student performance: A benchmarking, generalizability, and validity study of the use of teacher work samples. *Journal of Personnel Evaluation in Education.*

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching.* Alexandria, VA: Association for Supervision and Curriculum Development.

Hawley, W., & Valli, L. (1999). The essentials of effective professional development: A new consensus. In Linda Darling-Hammond and Gary Sykes (eds.), *Teaching and learning as a profession* (pp. 127-150). San Francisco: Jossey-Bass.

Idaho State Board of Education (2000). *Idaho standards for initial certification of professional school personnel.* Boise, ID: Idaho State Department of Education.

Marzano, R., & Kendall, J. (1998). *Awash in a sea of standards.* Aurora, CO: Mid-Continent Regional Educational Laboratory.

National Board for Professional Teaching Standards. (2001). *The effect of National Board Certification on teachers.* Washington, DC: National Board for Professional Teaching Standards.

National Commission on Teaching and America's Future. (1996). *What matters most: Teaching for America's future.* New York: National Commission on Teaching and America's Future.

National Council for Accreditation of Teacher Education. (2000). *NCATE 2000 Unit Standards.* Washington, DC: Author.

Popham, W. J. (1997). The moth and the flame: Student learning as a criterion of instructional competence. In J. Millman (Ed.). *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 264-274). Thousand Oakes, CA: Corwin Press.

Salzman, S. A., Denner, P. R., Bangert, A. W.& Harris, L. B. (2001, March). *Connecting Teacher Performance to the Learning of All Students: Ethical Dimensions of Shared Responsibility.* Symposium paper presented at the 53rd annual meeting of the American Association of Colleges for Teacher Education, Dallas, Texas. (ERIC Document Reproduction No. ED 451 182).

Salzman, S. A., & Denner, P. R. (2002, February). Teacher work samples: Evidence for teacher impact on student learning. In R. Pankratz (Chair), *The Renaissance partnership teacher work sample: A results-oriented performance assessment model with shared accountability.* Symposium conducted at the 54th annual meeting of the American Association of Colleges for Teacher Education, New York, New York.

Salzman, S. A., Denner, P. R., & Harris, L. B. (2002, February). *Teacher education outcomes measures: Special study survey.* Paper presented at the 54th annual meeting of the American Association of Colleges for Teacher Education, New York, New York.

Schalock, M. (1998). Accountability, student learning, and the preparation and licensure of teachers: Oregon's teacher work sample methodology. *Journal of Personnel Evaluation in Education, 12*, 269-285.

Schalock, M., Cowart, B., & Staebler, B. (1993). Teacher productivity revisited: Definition, theory, measurement, and application. *Journal of Personnel Evaluation in Education, 7*, 179-196.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Sykes, G. (1999). Teacher and student learning: Strengthening their connection. In Linda Darling-Hammond and Gary Sykes (eds.), *Teaching and learning as a profession* (pp. 151-179). San Francisco: Jossey-Bass.

Table 1

Number and Percent of Expert Raters Indicating Alignment Between TWS Guidelines, TWS Standards and TWS Scoring Rubric (N = 20)

Alignment Considerations	Degree of Alignment			
	Poor 1	Low 2	Moderate 3	High 4
Alignment of the TWS elements with the targeted TWS standards			4 20.0%	16 80.0%
Alignment of the TWS elements with the analytic scoring rubric			3 15.8%	16 84.2%
Alignment of the analytic scoring rubric with the targeted TWS standards			2 10.5%	17 89.5%

Table 2

Number and Percent of Expert Raters Indicating How Frequently They Would Expect a Teacher to Engage in the Teaching Behaviors Targeted by the TWS (N = 20)

Teaching Behaviors Targeted By Teacher Work Sample	Never	Yearly	Monthly	Weekly	Daily
Use information about the learning-teaching context and student individual differences to plan instruction and assessment.			1 5%	3 15%	16 80%
Set important, challenging, varied, and appropriate achievement targets.			4 20%	11 55%	5 25%
Use multiple assessment modes and approaches aligned with achievement targets to assess student learning before, during, and after instruction.			4 20%	6 30%	10 50%
Design instruction for specific achievement targets, student characteristics and needs, and learning contexts.			4 20%	4 20%	12 60%
Use assessment data to profile student learning, communicate information about student progress, and plan future instruction.		1 5%	4 20%	10 50%	5 25%
Reflect on his or her instruction and student learning in order to improve his or her teaching.				1 5%	19 95%

Table 3

Number and Percent of Expert Raters Indicating the Importance to Effective Teaching of the Teaching Behaviors Targeted by the TWS (N = 20)

Teaching Behaviors Targeted By Teacher Work Sample	Degree of Importance			
	Not at all Important 1	Somewhat Important 2	Important 3	Very Important 4
Use information about the learning-teaching context and student individual differences to plan instruction and assessment.			4 20%	16 80%
Set important, challenging, varied, and appropriate achievement targets.			4 20%	16 80%
Use multiple assessment modes and approaches aligned with achievement targets to assess student learning before, during, and after instruction.		1 5%	3 15%	16 80%
Design instruction for specific achievement targets, student characteristics and needs, and learning contexts.			5 25%	15 75%
Use assessment data to profile student learning, communicate information about student progress, and plan future instruction.			5 25%	15 75%
Reflect on his or her instruction and student learning in order to improve his or her teaching.			3 15%	17 85%

Table 4

Number and Percent of Expert Raters Indicating How Authentic the Tasks Required by the Teacher Work Sample Are to Success as a Classroom Teacher (N = 20)

Tasks Required By the Teacher Work Sample	Degree of Authenticity			
	Not at all Authentic 1	Somewhat Authentic 2	Authentic 3	Very Authentic 4
Analyze how school characteristics, classroom characteristics, and student characteristics impact instructional planning, delivery, and assessment.		3 15%	8 40%	9 45%
Establish important, challenging, varied, and appropriate achievement targets that clearly define what students are expected to know and be able to do as a result of an instructional sequence.		1 5%	4 20%	15 75%
Design an assessment plan to monitor student progress toward achievement targets that assess student performance before, during, and after instruction, including adaptations for students with special needs.		2 10%	7 35%	11 55%
Plan an instructional sequence, including at least six learning activities, aimed at specific achievement targets.		8 40%	3 15%	9 45%
Document delivery of an instructional sequence designed to facilitate accomplishment of specific achievement targets, including procedures and time lines, material and resources used, adaptations for students with special needs, and samples of student work that represent different levels of performance.		6 30%	8 40%	6 30%
Summarize student learning, including graphs or charts that profile student performance on pre-assessments and post-assessments, and disaggregate assessment data to analyze trends or differences in student learning.	5 25%	7 35%	5 25%	3 15%
Evaluate the effectiveness of an instructional sequence and reflect upon teaching practices and their effects on student learning, determining the extent to which achievement targets were met, actions to be taken next, the aspects of the teaching sequence that were especially successful, and how instruction might be done differently in the future.		4 20%	5 25%	11 55%

Table 5

Number and Percent of Expert Raters Indicating the Degree to Which the Tasks Required by the Teacher Work Sample Reflect and Represent the Targeted Standards (N = 20)

Tasks Required By the Teacher Work Sample	Degree of Representativeness			
	Not at all Representative 1	Somewhat Representative 2	Representative 3	Very Representative 4
Analyze how school characteristics, classroom characteristics, and student characteristics impact instructional planning, delivery, and assessment.		1 5%	10 50%	9 45%
Establish important, challenging, varied, and appropriate achievement targets that clearly define what students are expected to know and be able to do as a result of an instructional sequence.			7 35%	13 65%
Design an assessment plan to monitor student progress toward achievement targets that assess student performance before, during, and after instruction, including adaptations for students with special needs.			8 40%	12 60%
Plan an instructional sequence, including at least six learning activities, aimed at specific achievement targets.		1 5%	5 25%	14 70%
Document delivery of an instructional sequence designed to facilitate accomplishment of specific achievement targets, including procedures and time lines, material and resources used, adaptations for students with special needs, and samples of student work that represent different levels of performance.			7 35%	13 65%
Summarize student learning, including graphs or charts that profile student performance on pre-assessments and post-assessments, and disaggregate assessment data to analyze trends or differences in student learning.		1 5%	8 40%	11 55%
Evaluate the effectiveness of an instructional sequence and reflect upon teaching practices and their effects on student learning, determining the extent to which achievement targets were met, actions to be taken next, the aspects of the teaching sequence that were especially successful, and how instruction might be done differently in the future.			6 30%	14 70%

Table 6

Number and Percent of Expert Raters Indicating the Extent to Which the Tasks Required by the Teacher Work Sample Reflect the Idaho Core Teacher Standards (N = 20)

Idaho Core Teacher Standards	Not at all	Implicitly	Directly
1. The teacher understands the central concepts, tools of inquiry, and structures of the discipline(s) taught and creates learning experiences that make these aspects of subject matter meaningful to students.		8 40%	12 60%
2. The teacher understands how students learn and develop, and provides opportunities that support their intellectual, social, and personal development.		12 60%	8 40%
3. The teacher understands how students differ in their approaches to learning and creates instructional opportunities that are adapted to learners with diverse needs.		5 25%	15 75%
4. The teacher understands and uses a variety of instructional strategies to develop students' critical thinking, problem solving, and performance skills.		6 30%	14 70%
5. The teacher understands individual and group motivation and behavior and creates a learning environment that encourages positive social interaction, active engagement in learning, and self-motivation.	2 10%	11 55%	7 35%
6. The teacher uses a variety of communication techniques including verbal, nonverbal, and media to foster inquiry, collaboration, and supportive interaction in and beyond the classroom.	1 5%	11 55%	8 40%
7. The teacher plans and prepares instruction based upon knowledge of subject matter, students, the community, and curriculum goals.		2 10%	18 90%
8. The teacher understands, uses, and interprets formal and informal assessment strategies to evaluate and advance student performance and to determine program effectiveness.		3 15%	17 85%
9. The teacher is a reflective practitioner who demonstrates a commitment to professional standards and is continuously engaged in purposeful mastery of the art and science of teaching.		7 35%	13 65%
10. The teacher interacts in a professional, effective manner with colleagues, parents, and other members of the community to support students' learning and well-being.	3 15%	12 60%	5 25%

Table 7

Repeated Measures Analysis of Variance for Effect of Rater on Total Analytic Score Ratings

Source	*df*	*F* Set A	Set B
Rater	3	4.31*	6.87*
Residual	57	(1.53)	(3.77)

Note. Values enclosed in parentheses represent mean square errors. Set A = 20 teacher work samples chosen as exemplars across performance levels rated by the same 4 raters. Set B = another 20 work samples chosen at random across performance levels rated by 4 different raters.

*$p < .05$

Table 8

Estimates of Variance Components for the Person and Rater Facets Based on the Total Analytic Score Ratings of the Set A and Set B TWS (r = 4; n = 20)

Source	Estimated Variance Components	
	Set A	Set B
Person	1.896	3.27
Rater	.155	1.107
Residual	1.754	3.772

Table 9

Dependability Coefficient Estimates by Number of Raters for Set A and Set B TWS

	Dependability Coefficient Estimates	
Number of Raters	Set A	Set B
1 Rater	.50	.40
2 Raters	.66	.57
3 Raters	.75	.67
4 Raters	.80	.73
5 Raters	.83	.77

Table 10

Repeated Measures Analysis of Variance for the Effects of Rater and Occasion on the Total Analytic Score Ratings for Teacher Work Sample Set C and Set D

Source	*df*	*F* Set C	*F* Set D
Rater	2	2.058	7.619*
Error 1	18	(1.191)	(5.285)
Occasion	1	.010	.040
Error 2	9	(1.609)	(6.637)
Raters by Occasion	2	.060	.469
Error 3	18	(.276)	(1.848)

Note. Values enclosed in parentheses represent mean square errors. Set C = 20 teacher work samples rated by the same 3 raters developed by 10 teacher education candidates on 2 occasions using the analytic scoring method. Set D = another 20 teacher work samples rated by 3 different raters developed by 10 other teacher education candidates on the same 2 occasions using the analytic scoring method.

*$p < .05$

Table 11

Estimates of Variance Components for the Person and Rater Facets for the Total Analytic Score Ratings of TWS Set C and TWS Set D Averaged Across the Occasion Facet (r = 3; n = 10 per occasion)

Source	Estimated Variance Components	
	Set C	Set D
Person*	1.317	3.780
Rater*	.064	1.651
Residual*	.596	2.642

* Denotes the variance components are averaged across the fixed effect of occasion.

Table 12

Repeated Measures Analysis of Variance for the Effects of Rater and Occasion using the Holistic Scoring Method for the Set C and Set D Teacher Work Samples

Source	*df*	*F* Set C	*F* Set D
Rater	2	.340	12.765*
Error 1	18	(.589)	(.283)
Occasion	1	.150	.479
Error 2	9	(.446)	(.489)
Raters by Occasion	2	.200	.281
Error 3	18	(.107)	(.306)

Note. Values enclosed in parentheses represent mean square errors. Set C = 20 teacher work samples rated by the same 3 raters developed by 10 teacher education candidates on 2 occasions using the holistic scoring method. Set D = another 20 teacher work samples rated by 3 different raters developed by 10 other teacher education candidates on the same 2 occasions using the holistic scoring method.

*$p < .05$

Table 13

Estimates of Variance Components for the Person and Rater Facets for the Holistic Score Ratings of TWS Set C and TWS Set D Averaged Across the Occasion Facet (r = 3; n = 10 per occasion)

Source	Estimated Variance Components	
	Set C	Set D
Person*	.387	.065
Rater*	.004	.167
Residual*	.294	.141

* Denotes the variance components are averaged across the fixed effect of occasion.Table 12

Table 14

Mean Scoring Time for each of the TWS Sets by Scoring Method

	Analytic Scoring Method			Holistic Scoring Method		
TWS Set	*n*	*M*	*SD*	*n*	*M*	*SD*
Set A	77	14.3	5.9			
Set B	78	24.0	13.9			
Set C	39	22.7	7.5	60	27.9	18.5
Set D	39	34.0	14.9	58	14.7	6.67

Table 15

Estimates of Variance Components for the Person and Rater Facets for the Quality of Sources of Evidence Ratings of Set A TWS (n = 20)

Source	Estimated Variance Components
Person	.132
Rater	17.497
Residual	10.577

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: **Teacher work sample assessment: An accountability method that moves beyond teacher testing to the impact of teacher performance on student learning**	
Author(s): **Peter R. Denner, Stephanie A. Salzman, & Larry B. Harris**	
Corporate Source: **Idaho State University**	Publication Date: **2-25-02**

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1 X	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1	Level 2A	Level 2B
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature: Peter R. Denner	Printed Name/Position/Title: **Peter R. Denner/Professor**
Organization/Address: **Campus Box 8059, ISU Pocatello, ID 83209**	Telephone: **(208)282-4230** FAX:
	E-Mail Address: **dennpete@isu.edu** Date: **3-6-02**

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org

EFF-088 (Rev. 2/2001)