RESEARCH MEMORANDUM

THE INFLUENCE OF CHANGES IN ASSESSMENT DESIGN ON THE PSYCHOMETRIC QUALITY OF SCORES

Edward W. Wolfe
Drew H. Gitomer

ETS Educational Testing Service
Princeton, New Jersey

January 2000

The Influence of Changes in Assessment Design on the Psychometric Quality of Scores

Edward W. Wolfe

Michigan State University

Drew H. Gitomer

Educational Testing Service

Abstract

This article addresses the problem of improving the measurement quality of a complex performance assessment through principled assessment design. We describe the characteristics and measurement impact of steps taken to improve assessment exercise design along with modifications in assessor training materials and procedures between the 1995-1996 and the 1996-1997 administrations of the National Board for Professional Teaching Standards Early Childhood/Generalist examination. Specifically, we describe how the revision of this assessment resulted in increases in the inter-assessor agreement, internal consistency, and generalizability of scores. All indices we examined improved as a result of the revisions. The results suggest that previously observed limits on the measurement quality of performance assessments due to the relatively small number of items that contribute to an assessment score can be altered significantly through attention to assessment design and related scoring processes.

Key Words: National Board for Professional Teaching Standards, Performance Assessment, Test Construction, Teachers, Reliability

The Influence of Changes in Assessment Design on the Psychometric Quality of Scores

Principles for the design of performance assessments are very much in their infancy. As Linn and Baker (1996) point out about the development of performance assessment tasks:

> Far too often at this relatively early stage tasks are "created" and then rationalized rather than carefully and systematically designed. More interestingly, design processes can influence external validity criteria, that is, how performance-based assessments perform. (p.99)

The absence of principled design has led to assessments that have been challenged as not having the psychometric qualities to justify high-stakes decisions (e.g., Koretz, Stecher, Klein, & McCaffrey, 1994; Wainer & Thissen, 1994). Performance assessments have had much lower reliabilities than are typically observed for tests consisting of objectively scored items (e.g., multiple-choice). Of course, the demands of a performance assessment typically result in many fewer items, leading to reduced reliability estimates even if the covariation among items is similar to that found on conventional assessments. However, low covariance among performance assessment tasks may also be a problem. For example, Ruiz-Primo, Baxter, and Shavelson (1993) found relatively little consistency between how individuals performed on different items on a science performance assessment. Their conclusion was that a large number of performance items is necessary to meet acceptable standards of reliability for a high-stakes assessment.

In this article, we suggest that reliability can be influenced substantially by improving the design and scoring of performance assessments. This study explores whether indices of psychometric quality can be improved when design and scoring decisions are made with conscious attention to evidential issues raised by Messick (1989), Mislevy (1994) and Gitomer & Steinberg (1999). If we think of

assessment as a process for amassing evidence to support inferences about an individual, then all aspects of an assessment must be fashioned so as to provide evidence that is interpretable and coherent. An adequate performance assessment design must address the following two questions satisfactorily:

1. Does the information provided in a response (the answer) lead to interpretations by assessors that are both consistent and relevant to the intended purpose of the item?
2. Does the assessment, taken as a whole, provide coherent evidence that supports one or more target inferences consistent with the purpose of the assessment?

Focusing on these questions, we discuss progress that has been made in the design and development of a high-stakes assessment for the National Board for Professional Teaching Standards (NBPTS, 1997, 1998). These assessments consist of a relatively small number (10) of complex assessment tasks that yield a great deal of performance evidence within a task, but that are summarized by a single score per task. Thus, since the amount of information (number of scores) is low by assessment standards, it is imperative that the pieces of information provide the clearest evidence and most coherent inferences about an individual as possible. To satisfy this goal, we have placed great effort in the overall design of the assessments, particularly in the areas of item development and scoring processes. This article reviews how changes in these aspects of the assessment impact the psychometric quality of the resulting scores.

The National Board Assessments

The mission of the National Board for Professional Teaching Standards is to institute a national certification system that allows teachers to demonstrate accomplishment of high standards in the teaching profession and through which the teaching profession is enhanced. To this end, the National

Board has developed several certification examinations based on specific professional criteria. Teachers who volunteer to attempt to attain National Board certification are required to complete a series of performance-based assessments that ask teachers to include videotapes of classroom instruction, examples of instructional materials and student work, all with extensive written commentary, in addition to completing a set of essays in a timed assessment center setting. All assessments are based on NBPTS Standards written for teachers of specific content and students from specific age ranges. Standards exist in content areas as diverse as mathematics, visual arts, vocational education, English/Language arts, science, physical education, etc. The National Board breaks age ranges into four groups, Early Childhood (students aged 3-8), Middle Childhood (7-12), Early Adolescence (11-15), and Adolescence through Young Adulthood (14-18+). There are fewer content specific standards written for teachers of younger children, who tend to be generalists. Correspondingly, there are no generalist certificates for teachers of Adolescence through Young Adulthood students.

The National Board assessments have always consisted of two major components: the portfolio, and the assessment center. The portfolio exercises[1] focus on critical aspects of teaching (for example, student assessment), which should be part of the in-class practice of all accomplished teachers. The portfolio also has been used to assess teachers' accomplishment with respect to working with families and within the profession. Teachers receive the portfolio materials and specifications and have most of the school year in which to complete the required assessment exercises. Portfolio entries require candidates to write about their teaching and also include artifacts that are relevant to the exercise. Portfolio entries are highly contextualized. Candidates must explain, analyze, and justify their practice in terms of their actual teaching context and evidence of teaching practice that develops in that context.

The assessment center has played a different role, as it is designed to assess content and content pedagogical knowledge independent of a teacher's particular context. Conducted in a much more traditional assessment context, candidates provide extended written responses to a small set of prompts that requires one day of testing.

These assessments first appeared as field tests in 1993. Though the portfolio and assessment center remain the two central components, their design has undergone significant change, ranging from changes in the number of individual exercise scores to the very nature of the tasks. Whereas early assessments often asked teachers to describe their practice and/or philosophy of teaching in the abstract, all portfolio exercises now ask teachers to ground their discussion in tangible artifacts such as videotapes of lessons or student work samples. So, instead of asking teachers to elaborate on their philosophy of assessment, teachers are asked to provide examples of assessments that they use in their classrooms together with student work in response to the assessment tasks. Early portfolio entries, in an attempt to demonstrate face validity, often asked candidates to integrate an enormous amount of evidence covering, for example, lesson planning, classroom interaction, and assessment in a single entry. Despite good intentions, these types of entries turned out to be poor assessment tasks for they required too much interpretation and guesswork by both candidates and assessors. Candidates were required to organize and select from a tremendous amount of information to "make a case" that they believed assessors would find compelling. Whether or not candidates gauged assessors' preferences well had a major effect on their entry score, independent of whether or not the candidate's teaching was accomplished. From the assessors' perspective, the task of judging this huge amount of information, organized idiosyncratically by candidates, became a high inference task for assessors. Often times, they

were not judging the evidence directly, but inferring that individual candidates would have demonstrated certain abilities if they had only been asked to provide such a demonstration.

The assessment center tasks have changed dramatically as well. Many early assessment center prompts asked candidates to describe their approaches or philosophies to teaching content or particular students. Responses were difficult to judge in that candidates could provide very abstract treatises that told very little about their understanding of content or teaching—facility with jargon could serve a candidate quite well. More recent prompts ask candidates to respond to specific situations, such as the teaching of specific content, the analysis of a specific student's work, or the analysis of specific instructional resources, for example. Assessment center prompts also varied in the amount of time in which candidates were allowed to respond.

The changes from the earliest assessments to the more recent versions have been so dramatic that comparisons of their psychometric qualities would not tell us much—we would be comparing two different beasts. However, beginning in 1995-1996, the assessments began to stabilize. As shown in Table 1, all assessments have ten exercises that include six portfolio exercises—four classroom-based portfolio exercises and two documented accomplishment exercises—and four assessment center exercises. Classroom based portfolio exercises all are grounded in classroom artifacts, two based on videotapes of classroom discourse and two based on student work artifacts. The documented accomplishment exercises in the portfolio ask candidates to document their work outside the classroom, explaining both what the accomplishments are and why they are significant. One exercise asks for evidence of accomplishment in working with students' families and the community, while the second exercise asks for evidence of accomplishment in working with professional colleagues and organizations. The assessment center now consists of four 90-minute blocks for all certificates.

<< Insert Table 1 About Here >>

Though the structure has remained stable over the last several years, in 1996-1997 a great deal of emphasis was placed on improving the assessment in two ways. First, we attempted to improve the quality of evidence generated by candidates by modifying exercise instructions. Second, we tried to improve how evidence was considered by assessors by revamping assessor training methods, modifying scoring materials, and revising scoring processes. The purpose of this study is to describe these changes and then to examine changes in assessment quality brought about by these modifications in assessment and scoring design.

Assessment Design and Evidence Generation

Candidates for NBPTS assessment are asked to provide evidence about their accomplishment as teachers, making the best cases possible through completion of the assessment exercises. The assessment, particularly the portfolio, does not attempt to sample representative practice, but asks teachers to present their best examples of teaching. Candidates are encouraged to select classroom-based evidence from the better part of a school year. For such an assessment, we expect that candidates are indeed showing themselves as best they can. In order to make valid inferences about a teacher's level of accomplishment, assessors need to be as sure as possible that if a candidate provides evidence of teaching that is less than accomplished, that it is not because the candidate has misunderstood the requirements and expectations for an assessment exercise.

In performance assessment, a significant challenge is to reduce the number of assumptions and inferential leaps that an assessor must make in rendering judgments about a performance. Judgments should be made on evidence presented by the candidate, no more and no less. Assessors should not be

forced into assuming that a candidate could have shown some ability "if they had only been asked" or that "they probably could have done it had they picked a different class to show."

This is not to say, however, that assessors do not make any inferences. In fact, assessors are accomplished teachers themselves and they do make inferences based on their expertise as teachers. However, these inferences should be based only on the evidence presented, not on evidence that they assume might have been presented. For example, if a teacher were to see a classroom that had students asking questions of each other in a respectful manner, an assessor might make a reasoned judgment, based on his/her experience, that the candidate had spent significant effort establishing a learning climate in which such interchange was valued and modeled. This is a different inference than one, for example, in which the assessor assumes that a discussion would have occurred had the candidate not misunderstood the exercise directions.

A number of systematic changes have been implemented in the assessments to increase the likelihood that what candidates are asked to present and what assessors expect to see are aligned. Most of these changes are described in detail in the Technical Analysis Report developed by ETS (1999). We illustrate these changes by presenting portions of an actual exercise in Appendix A. To begin with, we have tried to reduce the guessing—candidates should not have to guess what assessors want to see and assessors should not have to guess what candidates might have said if given more precise directions. We have tried to reject the notion sometimes present in testing contexts that "the good ones will know what we're after."

The first major change instituted last year was the inclusion of the "How My Response will be Scored" section. This section is actually an approximation of the 4-level (highest level) of the rubric exercise. This section tells candidates exactly what assessors will value when their response is scored.

Note that the language in this section and the corresponding rubric refers to qualities that are sought in the response, rather than specific behaviors. The NBPTS assessments do not attempt to be prescriptive in terms of particular ways in which a teacher can be accomplished, but instead try to recognize that accomplished teaching can be realized with a variety of approaches.

A second addition to the entry directions in the portfolio is the "Making Good Choices" section. This section was written to help candidates make decisions that are likely to help them (and correspondingly protect them from making poor decisions) as they craft their entry. This section does not deal with the logistics of the entry, but rather with making and avoiding decisions that will support and hurt their entry, respectively. This section typically includes, as appropriate, suggestions for the selection of classes to videotape or students to follow, and for selecting instructional units and activities. For example, while most teachers of students of this age will have children engage in some type of drill and practice, such activities are probably not the best opportunity to showcase classroom discussion. This is not to say that discussion can not happen in such a context, but that in reviewing the work of previous candidates, such a choice is likely to make it more difficult to demonstrate accomplished practice.

The third change made was to add more structure to the questions that candidates responded to in their commentary for each entry. Earlier assessments tended to have fewer questions with less guidance about how to structure the response and allocate relative emphasis to different sections. In a sense, candidates were given a broad set of questions and asked to structure an essay addressing those questions.

Responses to these entries suggested that candidates might not be giving sufficient attention to some issues while overly attending to others. They might organize their response in ways that made it

more difficult for an assessor to locate evidence as well. In order to avoid having candidates make assumptions about how much to attend to each issue, the commentary is broken down into specific questions, with guidelines for page limits given as well. While still conducive to an integrated essay, the questions and questioning structure are designed to cue the candidate in how to organize the essay and how to attend to different issues with appropriate emphasis.

Scoring Design and Assessing Evidence

A great deal of work was done between the 1995-1996 and the 1996-1997 scoring sessions to improve the scoring design. Thompson (1998) reported how the scoring process was revised in 1996-1997. Here, we highlight some of the key changes that were implemented. Changes were designed to discipline the reading and interpreting of assessment evidence, ensuring that judgments remain governed by the rubric and standards only and grounded in the evidence presented. Training is designed to reduce, if not eliminate, the tendency for idiosyncratic considerations to be brought to bear on the judging of evidence, or for the possibility for going beyond the evidence and making judgments that require unsupported inferences.

Key changes in 1996-1997 included an increase in the number of benchmark and training samples the assessors were exposed to during training. These samples are used to provide illustrative images of different score points to assessors undergoing training. Rubrics and associated verbal descriptions are inherently limited. It is only by using actual examples that assessors could hone their judgments and learn the different ways in which scores at different levels could be achieved. Training samples were used to refine judgments by highlighting potentially ambiguous or distracting evidence in a candidate's responses. The additional use of these examples also resulted in an increase in the amount of time allocated to assessor training.

As Thompson (1998) notes, bias training was interleaved between the processing of these samples, also adding time to the training process. The bias training was a new process to enhance the likelihood of sound judgments grounded in the rubric and the reduction of judgments irrelevant to the rubric. Bias training raised issues of race, gender, and socioeconomic status of teachers and students. It also raised other issues that were not relevant to the rubrics but that could influence an assessor's judgment. For example, did an assessor have a preference for small group instruction or for teaching a particular concept in a certain way? While legitimate preferences in one's own classroom, these could be problematic biases during assessment. The purpose of the bias training was not to eliminate these preferences in their own teaching but to help assessors understand where and when such preferences could influence their judgment inappropriately and to refrain from doing so.

Other processes and structures were also put in place to refine and stabilize assessors' judgments. Among these was an explicit articulation of a model of teaching that underlies the NBPTS assessments. This architecture of teaching is an abstraction that serves to keep all assessors, across exercises and certificates, tied to a common framework for thinking about accomplished teaching.

Also put in place in 1996-1997 were guiding and bridge questions that served to structure the way in which assessors considered the evidence produced by candidates (Table 2). These questions were designed to keep assessors focused on the judgments they were required to make and reduced the possibility of assessors focusing on the less relevant or obscure aspects of a candidate's response. Note too, the changes in the scoring path. Whereas, the scoring path for the previous year was primarily procedural, this past year's document focused assessors much more on the analysis of evidence produced in the response.

<< Insert Table 2 About Here >>

Finally, the rubrics themselves have undergone significant change in focus and structure. Prior years tended to be more analytic, highlighting specific behaviors that might be observed at a score point. As discussed in the context of the "How my Response will be Scored" section, the current rubrics consciously avoid noting the presence or absence of specific behaviors at any score point. The problem with including specific behaviors in a holistic rubric is that an assessor may be at sea when the weight of evidence suggests one point on the scale, but an expected behavior for that score point is not observed (or vice versa). In such cases, assessors will often invent rules to deal with this conflicting information. Under the current scheme, assessors only weigh the preponderance of evidence regarding observed qualities of performance—they do not have to account for the presence or absence of specific acts.

Research Questions

Our purpose in this article is to evaluate the influence of the aforementioned changes in the National Board assessments on the psychometric quality of certification measures. More specifically, we address the following questions.

1) How do the assessment revisions influence inter-assessor agreement? Do the changes made in the instrument design make the search for evidence more consistent among assessors? Do changes in assessor training lead to more consistent judgments?
2) How do the assessment revisions influence inter-exercise consistency? Do attempts to reduce the introduction of bias in judgments and the articulation of a common framework for teaching lead to different patterns of consistency across assessment tasks?

3) How do the assessment revisions influence the generalizability of the measures? Taken as a whole, how do the changes in assessment design and scorer training influence the generalizability of the NBPTS assessments?

Method

Sample

To answer these questions, we compared measures derived from examinee responses to the 1995-1996 and the 1996-1997 National Board Early Childhood/Generalist certification examination. There were 234 examinees in 1995-1996 and 186 examinees in 1996-1997. As shown in Table 3, the demographic characteristics of the two samples varied only slightly. The 1996-1997 cohort was slightly more homogeneous with respect to geographic location and ethnicity but contained slightly more males than did the 1995-1996 cohort. In addition, the 1996-1997 cohort's ages were slightly more homogeneous than the previous year's cohort ($\underline{M}$ = 44, $\underline{SD}$ = 7.68 and $\underline{M}$ = 43, $\underline{SD}$ = 8.22, respectively).

<< Insert Table 3 About Here >>

Table 4 summarizes the professional characteristics of the two samples. From these figures, the two cohorts seem very similar with respect to the distributions of education level and teaching experience. There are slight differences in the other variables, however. For example, there were more teachers from rural districts and fewer from urban districts in 1996-1997. There are also differences between content certifications for the two samples. (NB, teachers could choose up to three areas, so differences in these figures could be the result of different rates of reporting between the two samples.) Slightly fewer teachers indicated each of the four major content areas among their areas of certification

and slightly more indicated other areas in the 1996-1997 cohort. There were also more non-responses to this question in the 1995-1996 cohort.

<< Insert Table 4 About Here >>

Instrument

Examinees respond to 10 exercises on the Early Childhood Generalist examination. Exercises from the 1995-1996 and 1996-1997 examinations are roughly parallel (recall Table 1), so comparisons for individual exercises and exercise types are possible. An assessor first assigned a whole number value of 1, 2, 3, or 4, and if appropriate, refined this judgment with either a plus (+) or a minus (–). A plus increased the whole number value by .25; a minus decreased the whole number value by .25. For example, a score of 2+ translated into a value of 2.25; a score of 4– translated into a value of 3.75. Because each whole number could be augmented with a plus or a minus, there were 12 possible score values ranging from .75 (1–) to 4.25 (4+). The rating scale is depicted as a number line with clustering of scores around the whole number values. This directs assessors to see the scale as composed of four distinct score "families," each with its own characteristics.

Assessors assigned a performance to a single score family, based on the preponderance of evidence in the response. Two assessors (nested within exercises) were randomly selected from the pool of assessors for a given exercise to score each response for each examinee. If the difference in the two assigned scores was 1.25 or less, then the two independent scores were averaged to yield an exercise score. If the difference between these two scores was 1.25 points or larger, then a third (more experienced) assessor gave a score which was then weighted with the other two scores to provide an exercise score. Hence, an exercise score was generated for each examinee by either averaging the two scores assigned by the assessors or by including a more highly weighted expert score in the case of

discrepancies. To determine a total assessment score for the individual, the ten exercise scores were weighted and then summed.

Analyses

Parallel analyses of the 1995-1996 and 1996-1997 Early Childhood Generalist data were performed, and results from each year were compared as a measure of the influence of instrument revision. Analyses focused on three factors: inter-assessor agreement, inter-exercise consistency, and generalizability.

Inter-Assessor Agreement

Inter-assessor agreement was evaluated in three ways for each data set. First, we examined the inter-assessor correlations for each data set. These correlations were computed via a Pearson Product Moment correlation from the randomly selected pairs of assessors for each examinee. That is, the inter-assessor correlations were computed on the first and second assessors for each examinee. Second, we examined the proportion of perfect agreement between the randomly selected pair of assessors in each data set and the proportion of perfect agreement corrected for chance agreement (i.e., coefficient κ). Third, we examined the resolution rates for each data set. That is, we identified the proportion of examinees for whom the randomly selected pair of assessors assigned scores that differed by 1.25 or more points.

Inter-Exercise Consistency

Inter-exercise consistency was evaluated in two ways for each data set. First, we examined the inter-exercise correlations for the composite exercise score for each data set. Second, we examined coefficient alpha for each data set. Coefficient alpha was computed using composite scores.

Generalizability

Generalizability was evaluated in two ways. First, we examined the reliability of scores from both data sets. To accomplish this, variance components were generated using a design in which examinees (*e*) are crossed with exercises (*i*), and assessors are nested within exercises (*r*) [i.e., a *e ´ (r : i)* design] (Brennan, 1992). We computed phi (i.e., based on absolute error, ϕ) and generalizability [i.e., based on relative error, $E(\rho^2)$] coefficients from these variance components. Second, we projected the number of additional assessors and the number of additional exercises that would be required to increase the reliability of 1995-1996 scores to the levels attained with the 1996-1997 scores.

Results

Inter-assessor Agreement

Table 5 shows the inter-assessor agreement indices for each exercise for the 1995-1996 and 1996-1997 data sets. As shown by these figures, there were large increases in the proportion of assigned scores in perfect agreement on the twelve-point rating scale. Most kappa indices increased by a similar magnitude. In addition, the resolution rate dropped in 1996-1997 on all of the exercises except one. On average, there was a drop in the resolution rate. Only one of the ten exercises was scored less consistently in 1996-1997—the portfolio entry concerning Engaging Students in Science Learning. Overall, the Documented Accomplishments entries showed the greatest improvements in inter-assessor agreement, with smaller improvements on the School Portfolio and the Assessment Center exercises. In addition, there was a fairly large increase in the average inter-assessor correlation across the assessment exercises. The largest average increase was observed for the Documented Accomplishment entries, with smaller increases for the Assessment Center and School Portfolio exercises.

<< Insert Table 5 About Here >>

Inter-Exercise Consistency

Table 6 shows the inter-exercise correlations for the 1995-1996 and 1996-1997 data sets. As shown by these figures, there was a modest increase in the average inter-exercise correlation across the assessment exercises. The average inter-exercise correlation for 1995-1996 was 0.29, while the average for 1996-1997 was 0.38. In general, the exercise scores were more consistent in 1996-1997 than in 1995-1996 (with a mean inter-exercise correlation increase of about 0.09). The inter-exercise correlation between the two Documented Accomplishment exercises (i.e., #5 and #6) showed a larger increase than did the correlations within the Assessment Center and the School Portfolio exercises. The increases in inter-exercise correlations between exercises of different formats (e.g., between exercise #1, a School Portfolio exercise, and exercise #4, a Documented Accomplishment exercise) were generally small. Coefficient alpha for the 1995-1996 and 1996-1997 data also indicated a modest increase in the internal consistency of the assessments (α = .83 and α = .88, respectively).

<< Insert Table 6 About Here >>

Generalizability

Table 7 shows the variance components and the ϕ and $E(\rho^2)$ coefficients for the 1995-1996 and the 1996-1997 data. Note that, for both data sets, the largest variance components are associated with error that is not taken into account by our generalizability study design with examinee and examinee-by-exercise effects accounting for the majority of the remaining variance. Exercise and assessor-within-exercise effects are small in both data sets. Also note that a fairly large decrease in error variance occurred between the 1995-1996 and 1996-1997 data and that this was associated with a somewhat large increase in examinee variance. As a result, there were substantial increases in both the ϕ (absolute) and $E(\rho^2)$ (relative) coefficients between 1995-1996 and 1996-1997.

<< Insert Table 7 About Here >>

What then are the practical consequences of the observed increase generalizability? We compared the observed increases of the generalizability of the 1996-1997 measures to increases that would be observed had we simply tried to increase the 1995-1996 reliability by adding exercises or assessors that were similar to those already included in the assessment. Table 8 projects the number of assessors and the number of exercises that would be required to increase the reliability of the 1995-1996 scores to the levels observed in the 1996-1997 data. As shown in the upper portion of Table 8, it would require over three times the number of current assessors to approximate the reliability levels attained through the revision efforts made for the 1996-1997 assessment. And, as shown in the lower portion of Table 8, one would need to nearly double the number of assessment tasks to obtain comparable reliabilities.

<< Insert Table 8 About Here >>

Discussion

Overall, the evidence provided here suggests that considerable increases in assessment quality can be obtained through revision of assessment materials and improvements in assessor training procedures. The 1995-1996 Early Childhood/Generalist assessment demonstrated reasonable levels of reliability, certainly in comparison with other complex performance assessments. The revisions that were made to the directions provided to examinees and the training materials and procedures provided to assessors resulted in non-trivial increases in those figures in 1996-1997. In addition, these increases in reliability were attained with minimal increases in costs. For example, revision of the examinee materials was done as part of the assessment development process, which would result in no additional

development costs. Revision of assessor training materials would result in only small additional development costs. Increasing the number of benchmarks that assessors review results in some increases in both development and scoring costs, but these increases are probably offset by a decrease in the number of examinee responses requiring adjudication. In fact, as our generalizability analyses show, obtaining comparable increases in reliability would require one to at least double the costs of administering or scoring the assessments.

These results suggest that findings of poor generalizability of performance assessments across task may lead some to conclude, erroneously, that the best way to improve the reliability of these assessments is to increase the number of tasks. Our results suggest that impressive gains can be obtained by careful consideration of the manner in which information is communicated to examinees and assessors. This implies that there is a significant burden on performance assessment developers to strengthen the quality and coherence of assessment tasks and overall instruments because financial constraints often necessitate a small number of tasks on a performance assessment instrument. Whereas traditional assessments can overcome less principled design characteristics by using many items, pre-testing, and not scoring identified items post-administration, those options generally are not available in complex performance assessments. Our data demonstrate that we can improve the technical quality of performance assessments by attending to the cognitive demands placed on those taking the assessments and those scoring them.

Footnotes

[1] NBPTS refers to portfolio tasks as entries and assessment center tasks as exercises. When considered collectively, NBPTS uses the term exercises, not items. This deliberate choice was made to highlight the distinction between the complex task demands of the NBPTS assessment tasks and the relatively limited task demands associated with traditional, short-answer assessments. In addition, those who judge the performances are designated as "assessors" rather than the commonly used term, "raters". We adopt the NBPTS terminology in this paper.

Author Note

The authors thank three anonymous reviewers and James Impara for thoughtful reviews of this manuscript. A previous version of this paper was presented at the annual meeting of the American Educational Research Association in San Diego, California, April, 1998. Correspondence concerning this article should be addressed to Edward W. Wolfe, 459 Erickson Hall, Michigan State University, East Lansing, MI, 48824-1034. Electronic mail may be sent via Internet to wolfee@msu.edu.

Appendix A

Sections from an Early Adolescence/English Language Arts Exercise

Analysis Of Student Writing

What Is the Nature Of This Entry?

In this entry you will demonstrate how you teach and analyze student writing. The entry asks you to submit **student writing** for three students and for each piece of student writing a **Written Commentary** about the goals for your teaching, the teaching context or assignment that led to the writing, and an analysis of each student as a developing writer. It also asks you to explain how you assessed the writing and presented feedback to the student, and to tell how you used this writing to build further instruction. Finally, the entry asks you to reflect on how the three entries, taken together, are indicative of your teaching of writing.

What Do I Need To Do?

Submit a **folder for each of three students** containing a **Student Work Caption Sheet** with answers to the three topics, **the student's piece of writing** with work that shows the student's writing process and any written feedback the student received, and a **Written Commentary** about the writing; and a brief **Reflective Essay** explaining how the three students' writing and related work and the context that shaped them are indicative of your practice as a teacher. (See **Making Good Choices** for more detail.) **Failure to submit any of these materials will make your response to this entry unscorable.**

For this entry you will

- Select **three students** who represent different kinds of challenges to you.

- Submit **a folder for each student** containing a **Student Work Caption Sheet**, a **piece of the student's writing** and other related work by the student that shows the writing process that the student used and feedback the student received, and a **three-page Written Commentary**. (See **Written Commentary** for more detail.)

- Submit **one two-page Reflective Essay** commenting on the work of the three students and how their work and the context are indicative of your practice. (See **Written Commentary** for more detail.)

Making Good Choices

You have two important choices to make for this entry. First, you will need to select **three students** to feature. It is important to choose students whose writing gives you an opportunity to discuss your

practice. For this reason, the best-performing students in the class may not be the best choices for this entry. The focus is on your practice, not on the level of student performance. Second, you must choose **three pieces of student writing – one for each student** – that, taken together, are indicative of your approach to teaching writing.

To prepare for this entry, over a period of time you will want to select several (at least six) students as potential cases and collect or make copies of written work for each of them, including all drafts and other work showing the writing process that they used. These students might be members of the same class or might be drawn from several different classes that you teach. They must all be ages 11 through 15 and none of them should be among the students chosen for the *Analysis of Student Response to Literature* entry. As you collect the work, record or take notes on your reasons for selecting that particular student and his/her work, and details that might be helpful in completing your analysis – for example, your learning goals, what came before and after the assignment, steps in the student's writing process, how you responded, how you built on your assessment of the student's writing.

After reviewing the work you have collected for each of these students, choose the three students whose work you will submit and about whom you will write by reviewing once again the entire *Analysis of Student Writing* materials. Choose three students who represent a range of kinds of writers and challenges to you as a teacher.

Then, for each of the three students, select one piece of writing, with the drafts and other student work that shows the writing process that the student used. The student writing you select can take many forms. Choose pieces of writing that illustrate different challenges, problems, or topics in the teaching of writing, and that illustrate your approaches to teaching writing. Be certain to select pieces of writing that are substantial enough to support a discussion of the kind outlined in the **Written Commentary** section below.

Also select any materials that explain what the assignment was or the conditions under which the assignment was done. Attach copies of these materials, if appropriate, to the **Student Work Caption Sheet**. (See **Student Work Caption Sheet/Student Response** below.)

♦ **Written Commentary**

Your **Written Commentary** must address the following elements and be organized into sections with the headings that appear in boldface below. Consistent headings will help assessors locate the required information more easily. The entire **Written Commentary** must be no longer than **11 typed pages.** (See Format Specifications below for more detail.)

Part I. Analysis of individual students: For each of the three students, submit a folder (marked for Student A, B, or C) containing the following:

1. **A Student Work Caption Sheet** that provides full responses to the three topics listed, with the materials described in your response to the third topic attached, if appropriate.

2. **The piece of student writing**

3. **A Written Commentary of no more than three pages that addresses each of the following. Please label each section with the heading that appears in boldface:**

   Instructional Context: In this section, address the following questions:

   - *What was the instructional sequence that led to the piece of student writing*?
   - *What prompted this piece of writing?*
   - *What were your instructional goals?*

   The Student: In this section, address the following question:

   - *What about the student (background, skills, and interests) helps explain this piece of writing?*

   The Student's Writing: In this section, address the following questions:

   - *What do you see as the special, defining characteristics of the writing?*
   - *What does it suggest about the student's development and accomplishment as a writer?*
   - *How did you assess this writing?*
   - *How did you present assessment feedback to the student?*

   Planning:
   - *In light of this student writing, what did you do next to build on what the student accomplished?*

   [*Total page length for the Written Commentary for each student must not be longer than 3 pages.]*

Part II: Reflective Essay: Write **an essay of no more than two pages** that addresses the following question:

- *How do these **three students' writings, considered together**, and the **teaching context** that shaped them demonstrate your goals and approaches to the teaching of writing and the challenges you face as a teacher of writing.* Use the three students' work you have submitted to illustrate your discussion.

How Will My Response Be Scored?

The following standards for accomplished Early Adolescence/English Language Arts practice constitute the criteria that will be applied to score your response to this entry. It is strongly recommended that you review these standards before beginning and periodically as you prepare your entry. This entry will be evaluated with respect to the following standards: **Standard I: Knowledge of Students; Standard II: Curricular Choices; Standard V: Instructional Resources; Standard VI: Reading; Standard XI: Assessment; and Standard XII: Self-Reflection.**

The response shows clear and consistent knowledge of individual students through descriptions of their backgrounds and skills as writers.

The response shows clear and consistent evidence of thoughtful analyses of student texts to understand the individual growth and development of student writers.

The response shows clear and consistent evidence of the ability to establish attainable and worthwhile learning goals and to make curricular choices and instructional resources designed to enable student writers to achieve those goals.

The response shows clear and consistent evidence of an understanding that writing is a complex, recursive thinking process and that writers vary widely in how they orchestrate the creative process, and of the ability to establish an instructional context for writing that encourages students' active exploration of individual writing processes.

The response shows clear and consistent evidence of the use of formal and/or informal assessment methods to monitor student progress, to encourage student self-assessment, and to plan instruction.

The response shows clear and consistent evidence of insightful and well-informed analyses of the teacher's classroom practices including a clear rationale for why those practices are appropriate for the students.

References

Brennan, R.L. (1992). Elements of Generalizability Theory. Iowa City, IA: ACT.

ETS (1999). National Board for Professional Teaching Standards: Technical Analysis Report 1996-1997 Administration. Princeton, NJ: Author.

Gitomer, D.H. & Steinberg, L.S. (1999). Representational issues in assessment design. In I.E. Sigel (Ed.) Development of mental representation: Theories and applications (pp. 351-369). Hillsdale, NJ: Lawrence Erlbaum.

Koretz, D.M., Stecher, B.M.,. Klein, S.P., & McCaffrey, D. (1994). The Vermont Portfolio Assessment Program: Findings and Implications. Education Measurement: Issues and Practice, 13, 5-16.

Linn, R. L. & Baker, E. L. (1996). Can performance-based student assessments be psychometrically sound? In J. B. Baron and D. P. Wolf (Eds.), Performance-based student assessment: Challenges and possibilities, Ninety-fifth Yearbook of the National Society for the Study of Education. Chicago: University of Chicago Press.

Mislevy, R.J. (1994). Evidence and inference in educational assessment. Psychometrika, 59, 439-483.

National Board for Professional Teaching Standards (1997). Performance assessment for NBPTS certification. Detroit: Author.

National Board for Professional Teaching Standards (1998). Performance assessment for NBPTS certification. Detroit: Author.

Ruiz-Primo, M.A., Baxter, G.P., & Shavelson, R.J. (1993). On the stability of performance assessments. Journal of Educational Measurement, 30, 41-53.

Thompson, M. (1998) Data quality as a function of different scoring models. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Wainer, H. & Thissen, D. (1994). On examinee choice in educational testing. Review of Educational Research; 64, 159-95.

Table 1

Assessment Structure for the 1995-1996 and 1996-1997 Assessments

Exercise	Format	Topic
1	School Portfolio	Classroom Community
2	School Portfolio	Teaching & Learning
3	School Portfolio	Engaging Students in Science Learning
4	Documented Accomplishments	Working with Families
5	School Portfolio	Literacy Development
6	Documented Accomplishments	Professional Collaborations
7	Assessment Center	Work Sample
8	Assessment Center	Curriculum
9	Assessment Center	Assessment
10	Assessment Center	Observing Children

Table 2

Bridge Questions

Purpose	Questions
To help assessors see different parts of the evidence	Are the goals of the lesson worthwhile and appropriate, even if they are not goals I would choose for my students?
	Is the teacher demonstrating knowledge of his or her students, as individuals or as a developmental or social group, even if the teacher's approach is different from one I would take?
	Is the teacher showing command of the content, making connections, even if they are not the connections I would make?
	Are students engaged in the lesson, even if it's not in a way I am used to?
	Is the teacher showing respect for all students, even if the teacher's style is different from mine?
	If there is something troubling to you about the teacher's choices (content, style, classroom organization, material), is there a plausible and professionally acceptable explanation that would explain why s/he made the choices s/he made?
To help assessors identify the underlying architecture of the performance	What is the underlying structure of this performance? What is going on beneath the surface features (e.g., level of resources in the classroom, teacher's and students' accents and appearance, noise level, writing ability demonstrated in a response)?
	As you begin to formulate a hypothesis about the accomplishment demonstrated in this performance, can you construct a counter-hypothesis that is also rooted in the evidence and the rubric?

Table 3

Demographic Characteristics for the 1995-1996 and 1996-1997 Assessments

	1995-1996	1996-1997
Variable	%	%
Geographic Location		
East	41.5	53.2
Central	47.4	40.8
West	11.1	5.9
Gender		
Female	99.2	97.7
Male	0.9	2.2
Ethnicity		
White	80.8	85.5
African American	12.0	8.1
Hispanic	5.6	3.8
Other / Blank	1.7	2.6
Age		
≤ 29	6.0	7.5
30 – 39	26.5	26.9
40 – 49	44.0	47.3
50 – 59	21.8	16.7
≥ 60	1.7	1.6

Note: Cell totals may not equal 100 because of rounding error.
n = 234 and 186 for 1995-1996 and 1996-1997, respectively.

Table 4

Professional Characteristics for the 1995-1996 and 1996-1997 Assessments

	1995-1996	1996-1997
Variable	%	%
District		
Urban	48.7	36.0
Rural	22.7	35.0
Suburban	28.2	27.4
Degree		
BA	27.8	28.5
MA	70.1	69.9
Ph.D.	2.1	1.1
Subject		
English Language Arts	87.6	80.1
Mathematic	83.3	79.0
Science	62.0	63.4
Social Studies / History	21.4	18.3
Other	12.9	15.1
Non Indicated	8.6	6.5
Years Teaching		
≤ 9	32.5	34.4
10 – 19	40.6	43.0
20 – 29	24.4	21.0
≥ 30	2.6	1.6

Note: Cell totals may not equal 100 because of rounding error, omits, or multiple answers.
n = 234 and 186 for 1995-1996 and 1996-1997, respectively.

Table 5

Agreement Rates for the 1995-1996 and 1996-1997 Assessments

		1995-1996				1996-1997			
Exercise	Format	Perfect	κ	Resolved	r	Perfect	κ	Resolved	r
1	SP	.16	.06	.15	.40	.22	.12	.06	.60
2	SP	.17	.07	05	.60	.26	.16	.06	.63
3	SP	.15	.05	.16	.35	.11	-.02	.10	.33
4	DA	.17	.08	.13	.54	.25	.14	.01	.74
5	SP	.23	.13	.08	.58	.27	.18	.05	.68
6	DA	.17	.06	.11	.42	.30	.19	.05	.69
7	AC	.21	.10	.07	.45	.23	.12	.04	.61
8	AC	.29	.16	.07	.40	.32	.19	.01	.63
9	AC	.18	.07	.13	.30	.26	.13	.06	.51
10	AC	.17	.08	.13	.50	.20	.11	.09	.56
Average		.19	.09	.11	.46	.24	.13	.05	.60

Note. "Perfect" refers to the proportion of exercises that were assigned the exact same rating on the 12-point scale by the two assessors. κ (kappa) is the proportion of perfect agreement corrected for chance agreement. Two independent scores were "resolved" if the difference between those scores was greater than 1.25. r is the Pearson Product Moment correlation between two raters for each exercise. SP=School Portfolio, DA=Documented Accomplishment, AC=Assessment Center.

Table 6

Inter-Exercise Correlations for the 1995-1996 and 1996-1997 Assessments

1995-1996

Exercise (Format)	1 (SP)	2 (SP)	3 (SP)	4 (DA)	5 (SP)	6 (DA)	7 (AC)	8 (AC)	9 (AC)	10 (AC)
1 (SP)	--	.59	.42	.61	.56	.42	.32	.26	.37	.38
2 (SP)	0.40	--	.44	.50	.45	.39	.25	.26	.26	.40
3 (SP)	0.33	0.33	--	.41	.35	.29	.24	.18	.19	.25
4 (DA)	0.37	0.39	0.23	--	.53	.57	.32	.34	.28	.32
5 (SP)	0.41	0.41	0.32	0.47	--	.34	.34	.37	.41	.40
6 (DA)	0.31	0.29	0.17	0.31	0.45	--	.33	.31	.32	.31
7 (AC)	0.17	0.27	0.18	0.25	0.30	0.20	--	.35	.52	.47
8 (AC)	0.32	0.23	0.10	0.12	0.15	0.16	0.25	--	.43	.40
9 (AC)	0.33	0.28	0.04	0.27	0.42	0.35	0.25	0.22	--	.48
10 (AC)	0.37	0.37	0.15	0.31	0.33	0.33	0.37	0.31	0.30	--

Note. Upper off-diagonal entries refer to inter-exercise correlations for the 1996-1997 data, and lower off-diagonal entries refer to the 1995-1996 data. The mean inter-exercise correlation for 1995-1996 = 0.29. The mean inter-exercise correlation for 1996-1997 = 0.38. SP=School Portfolio, DA=Documented Accomplishment, AC=Assessment Center.

Table 7

Variance Components for the 1995-1996 and 1996-1997 Assessments

Facet	1995-1996 Variance Component	1996-1997 Variance Component
Examinee	0.14 (18%)	0.21 (28%)
Exercise	0.06 (8%)	0.06 (8%)
Assessor : Exercise	0.04 (5%)	0.02 (3%)
Examinee × Exercise	0.20 (25%)	0.19 (25%)
Error	0.35 (44%)	0.27 (36%)
ϕ	.75	.84
$E(\rho^2)$	.78	.87

Table 8

Decision Studies for the 1995-1996 Assessment – The Increase in Assessors or Exercises Required to Attain Generalizability Estimates Observed for the 1996-1997 Assessment

Coefficient	Assessors				
	5	10	15	20	25
ϕ	.80	.82	.83	.83	.83
$E(\rho^2)$	.83	.85	.86	.86	.86

Coefficient	Exercises				
	13	15	16	17	18
ϕ	.79	.82	.83	.83	.84
$E(\rho^2)$	.82	.84	.85	.86	.87

Note. These values should be compared to the 1996-1997 values of ϕ=.84 and $E(\rho^2)$=.87.