APPLIED MEASUREMENT IN EDUCATION, *14*(1), 91–107

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

Edward W. Wolfe
Department of Measurement and Quantitative Methods
Michigan State University

Drew H. Gitomer
ETS
Princeton, New Jersey

In this article, we address 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 contributed to increases in the interassessor agreement, internal consistency, and generalizability of scores. All indexes we examined improved after 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 may be altered significantly through attention to assessment design and related scoring processes.

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

Requests for reprints should be sent to Edward W. Wolfe, 459 Erickson Hall, Michigan State University, East Lansing, MI 48824–1034. E-mail: wolfee@msu.edu

> 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 typically are 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 also might 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 indexes 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 and 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). 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 are summarized by a single score per task. Thus, because 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. We review how changes in these aspects of the assessment impact the psychometric quality of the resulting scores.

NBPTS

The mission of the National Board of Assessments (NBPTS) 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 NBPTS has developed several certification examinations that require candidates to complete performance-based assessments including videotapes of classroom instruction, examples of instructional materials and student work with written commentary, and writing a set of essays in a timed assessment center setting. All assessments are based on NBPTS standards written for teachers of a variety of content areas and students from specific age ranges.

The NBPTS assessments always have consisted of two major components: the portfolio and the assessment center. The portfolio exercises[1] focus on critical aspects of teaching that should be part of the in-class practice of all accomplished teachers as well as 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 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 1 day of testing.

Although the portfolio and assessment center have always been the two central components of the NBPTS assessments, their design has undergone several significant change since their inception in 1993. First, whereas early assessments often asked teachers to describe their practice or philosophy of teaching in the abstract,

[1]The 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 article.

portfolio exercises now ask teachers to ground their discussion in tangible artifacts such as videotapes of lessons or student work samples. Second, early portfolio entries often asked candidates to integrate an enormous amount of evidence in a single entry (e.g., lesson planning, classroom interaction, and assessment). These types of entries turned out to be poor assessment tasks because they required too much interpretation and guesswork by both candidates and assessors. Now the portfolio entries are more focused in the requested information. Third, 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. 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. Finally, assessment center prompts used to vary in the amount of time in which candidates were allowed to respond. Now comparable time is provided for each response.

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 10 exercises that include 6 portfolio exercises—4 classroom-based portfolio exercises and 2 documented accomplishment exercises—and 4 assessment center exercises. Classroom-based portfolio exercises all are grounded in classroom artifacts, 2 based on videotapes of classroom discourse and 2 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, whereas the second exercise asks for evidence of accomplishment in working

TABLE 1
Assessment Structure for the 1995–1996 and 1996–1997 Assessments

Exercise	*Format*	*Topic*
1	School portfolio	Classroom Community
2	School portfolio	Teaching and 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

with professional colleagues and organizations. The assessment center now consists of four 90-min blocks for all certificates.

Although 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 examine changes in assessment quality brought about by these modifications in assessment and scoring design.

Assessment Design and Evidence Generation

Candidates for NBPTS certification 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 one school year. For such an assessment, we expect that candidates are indeed showing themselves as best they can. 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 or 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 summarize those changes here. To begin with, we 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 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 was the inclusion of the "How My Response Will Be Scored" section. This section is actually an approximation of the four-level (highest level) of the rubric exercise. It tells candidates exactly what assessors will value when their response is scored. 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. The text 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, whereas 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 cannot 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. 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. Although 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) noted, 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? Although 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.

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 is 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 she or he made the choices?
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?

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 1996–1996 year was primarily procedural, the 1996–1997 scoring path document focused assessors much more on the analysis of evidence produced in the response.

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 might 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 often will invent rules to deal with this conflicting information. Under this

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 NBPTS assessments on the psychometric quality of certification measures. More specifically, we address the following questions.

1. How do the assessment revisions influence interassessor 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 interexercise 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 NBPTS 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 men than did the 1995–1996 cohort. In addition, the 1996–1997 cohort's ages were slightly more homogeneous than the previous year's cohort (M = 44, SD = 7.68 and M = 43, SD = 8.22, respectively).

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 con-

tent certifications for the two samples. (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 nonresponses to this question in the 1995–1996 cohort.

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

TABLE 3
Demographic Characteristics for the 1995–1996 and 1996–1997 Assessments

Variable	*1995–1996*[a]	*1996–1997*[b]
Geographic location		
East	41.5	53.2
Central	47.4	40.8
West	11.1	5.9
Gender		
Women	99.2	97.7
Men	0.9	97.7
Ethnicity		
White	80.8	85.5
African American	12.0	8.1
Hispanic	5.6	3.8
Other/blank	1.7	2.6
Age		
Less than 29	6.0	7.5
30 to 39	26.5	26.9
40 to 49	44.0	47.3
50 to 59	21.8	16.7
Less than 60	1.7	1.6

Note. Cell totals may not equal 100 because of rounding error. Given in percentages.
[a]$n = 234$. [b]$n = 186$.

TABLE 4
Professional Characteristics for the 1995–1996 and 1996–1997 Assessments

Variable	*1995–1996*[a]	*1996–1997*[b]
District		
Urban	48.7	36.0
Rural	22.7	35.0
Suburban	28.2	27.4
Degree		
B.A.	27.8	28.5
M.A.	70.1	69.9
Ph.D.	2.1	1.1
Subject		
English language arts	87.6	80.1
Mathematics	83.3	79.0
Science	62.0	63.4
Social studies/history	21.4	18.3
Other	12.9	15.1
Not indicated	8.6	6.5
Years teaching		
Less than 9	32.5	34.4
10 to 19	40.6	43.0
20 to 29	24.4	21.0
More than 30	2.6	1.6

Note. Cell totals may not equal 100 because of rounding error. Given in percentages.
[a]$n = 234$. [b]$n = 186$.

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 more than 1.25 points, then a third (more experienced) assessor gave a score that 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 10 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: interassessor agreement, interexercise consistency, and generalizability.

Interassessor agreement. Interassessor agreement was evaluated in three ways for each data set. First, we examined the interassessor 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 interassessor 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 more than 1.25.

Interexercise consistency. Interexercise consistency was evaluated in two ways for each data set. First, we examined the interexercise correlations for the composite exercise score for each data set. Second, we examined coefficient α for each data set. Coefficient α 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 $e \times (r{:}i)$ design] (Brennan, 1992). We computed ϕ (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

Interassessor Agreement

Table 5 shows the interassessor agreement indexes for each exercise for the 1995–1996 and 1996–1997 data sets. As shown by these figures, there were large

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 = the proportion of exercises that were assigned the exact same rating on the 12-point scale by the two assessors; κ = the proportion of perfect agreement corrected for chance agreement; resolved = when two independent scores were resolved if the difference between those scores was greater than 1.25; *r* = Pearson Product Moment correlation between two raters for each exercise; SP = School Portfolio; DA = Documented Accomplishment; AC = Assessment Center.

increases in the proportion of assigned scores in perfect agreement on the 12-point rating scale. Most κ indexes 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 1 of the 10 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 interassessor agreement, with smaller improvements on the School Portfolio and the Assessment Center exercises. In addition, there was a fairly large increase in the average interassessor 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.

Interexercise Consistency

Table 6 shows the interexercise correlations for the 1995–1996 and 1996–1997 data sets. As shown by these figures, there was a modest increase in the average interexercise correlation across the assessment exercises. The average interexercise 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 interexercise correlation increase of about 0.09). The interexercise 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 interexercise 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 α for the 1995–1996 and 1996–1997 data also indicated a modest increase in the internal consistency of the assessments ($\alpha = .83$ and $\alpha = .88$, respectively).

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. In addition, 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.

TABLE 6
Interexercise Correlations for the 1995–1996 and 1996–1997 Assessments

Exercise	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
(Format)	*(SP)*	*(SP)*	*(SP)*	*(DA)*	*(SP)*	*(DA)*	*(AC)*	*(AC)*	*(AC)*	*(AC)*
1 (SP)	—	.59	.42	.61	.56	.42	.32	.26	.37	.38
2 (SP)	.40	—	.44	.50	.45	.39	.25	.26	.26	.40
3 (SP)	.33	.33	—	.41	.35	.29	.24	.18	.19	.25
4 (DA)	.37	.39	.23	—	.53	.57	.32	.34	.28	.32
5 (SP)	.41	.41	.32	.47	—	.34	.34	.37	.41	.40
6 (DA)	.31	.29	.17	.31	.45	—	.33	.31	.32	.31
7 (AC)	.17	.27	.18	.25	.30	.20	—	.35	.52	.47
8 (AC)	.32	.23	.10	.12	.15	.16	.25	—	.43	.40
9 (AC)	.33	.28	.04	.27	.42	.35	.25	.22	—	.48
10 (AC)	.37	.37	.15	.31	.33	.33	.37	.31	.30	—

Note. Upper off-diagonal entries refer to interexercise correlations for the 1996–1997 data, and lower off-diagonal entries refer to the 1995–1996 data. The mean interexercise correlation for 1995–1996 = .29. The mean interexercise correlation for 1996–1997 = .38. SP = School Portfolio; DA = Documented Accomplishment; AC = Assessment Center.

TABLE 7
Variance Components for the 1995–1996 and 1996–1997 Assessments

	1995–1996		*1996–1997*	
Facet	*No.*	*%*	*No.*	*%*
Examinee	.14	18	.21	28
Exercise	.06	8	.06	8
Assessor: exercise	.04	5	.02	3
Examinee × Exercise	.20	25	.19	25
Error	.35	44	.27	36
ϕ	.75		.84	
$E(\rho^2)$	.78		.87	

TABLE 8
Decision Studies for the 1995–1996 Assessment: Increase in Assessors or Exercises Required to Attain Generalizability Estimates Observed for the 1996–1997 Assessment

	Assessors					*Exercises*				
Coefficient	*5*	*10*	*15*	*20*	*25*	*13*	*15*	*16*	*17*	*18*
ϕ	.80	.82	.83	.83	.83	.79	.82	.83	.83	.84
$E(\rho^2)$	.83	.85	.86	.86	.86	.82	.84	.85	.86	.87

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

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 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.

DISCUSSION

To summarize, we see a fairly substantial increase in the reliability of scores between the 1995–1996 and the 1996–1997 administrations of the NBPTS Early

Childhood/Generalist assessment. Because of the variety of uncontrolled variables in our study, we acknowledge that there are at two least uncontrolled factors that might have produced these increases. First, from the generalizability study data, one might conclude that the increase in reliability was due to an increase in the variance of the sample. However, we believe that this is unlikely, given the demographic and professional characteristics of the two samples compared here (Tables 3 and 4) because the samples are less heterogeneous on several of these variables in the 1996–1997 sample (e.g., geographic location, ethnicity, age, years of teaching).

Second, it is possible that these increases are due to learning on the part of candidates or assessors who took part in the assessment process in both years. Such learning could result in candidate materials that are more consistent with the scoring guidelines or assessors who are better able to agree. It is unlikely that the observed changes could be attributed to learning on the part of repeat candidates. Out of curiosity, we merged the two data sets, matching on state and birth year. Of the 420 combined records, only 18 of the cases matched on these variables (4%). With respect to assessors, about 13% of the assessors for the 1996–1997 scoring returned from the 1995–1996 session.

In light of the evidence, we believe that the most reasonable explanation for the increase in reliability between 1995–1996 and 1996–1997 lies in the revisions of the assessment materials and improvements in assessor training procedures. The 1995–1996 Early Childhood/Generalist assessment demonstrated reasonable levels of reliability, but the figures from 1996–1997 are clearly better. What is important is that 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 tasks 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, pretesting, and

not scoring identified items postadministration, 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.

ACKNOWLEDGMENTS

A previous version of this article was presented at the annual meeting of the American Educational Research Association in San Diego, California, April, 1998. We thank three anonymous reviewers and James Impara for thoughtful reviews of this manuscript.

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). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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*(3), 5–16.

Linn, R. L., & Baker, E. L. (1996). Can performance-based student assessments be psychometrically sound? In J. B. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities, ninety-fifth yearbook of the National Society for the Study of Education* (pp. 84–103). Chicago: University of Chicago Press.

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

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, April). *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.