

## Review of the Psychometric Quality of the National Board Assessments

Teresa L. Russell  
Dan Putka  
Shonna Waters

*Prepared for:*

The National Academies  
National Academy of Sciences  
National Research Council  
Committee on the Evaluation of Teacher Certification by the National Board of  
Professional Teaching Standards  
500 Fifth Street, NW  
Washington, DC 20001

Grant No: DBASSE-6062-06-001

15 October 2007

## Review of the Psychometric Quality of the National Board Assessments

Teresa L. Russell  
Dan Putka  
Shonna Waters

*Prepared for:*

The National Academies  
National Academy of Sciences  
National Research Council  
Committee on the Evaluation of Teacher Certification by the National Board of  
Professional Teaching Standards  
500 Fifth Street, NW  
Washington, DC 20001

Grant No: DBASSE-6062-06-001

15 October 2007

**REVIEW OF THE PSYCHOMETRIC QUALITY OF THE NATIONAL BOARD  
ASSESSMENTS**

**Table of Contents**

<b>Purpose.....</b>	<b>1</b>
<b>Approach.....</b>	<b>1</b>
<b>Development, Administration and Scoring of NBPTS Assessments .....</b>	<b>2</b>
Development of Content Standards .....	2
Standards Development Committees .....	3
Drafting the Standards.....	4
Getting Input on the Standards .....	4
Publishing and Updating the Standards.....	4
Development of Assessment Exercises.....	4
Teacher Assessment Development Teams .....	5
Portfolios.....	6
Assessment Center Exercises.....	6
Pilot Testing and Formative Scoring of Assessments.....	7
Administration of Assessments .....	8
Portfolios.....	8
Assessment Center Exercises.....	8
Benchmarking.....	9
Scoring of Assessments.....	10
Small Sample Scoring .....	10
Content Validity Studies .....	10
<b>Technical Characteristics: Reliability, Fairness, and Validity.....</b>	<b>11</b>
Reliability .....	11
Potential Sources of Measurement Error.....	12
Assessors and Exercises .....	12
Estimating Reliability.....	13
Estimating Decision Accuracy.....	20
Fairness.....	22
Subgroup Differences and Disparate (Adverse) Impact.....	22
NBPTS Research on Disparate (Adverse) Impact .....	24
<b>Conclusions .....</b>	<b>25</b>
<b>References .....</b>	<b>26</b>

## List of Tables

Table 1. General Assessment Development Guidelines .....	5
Table 2. Comparison of Methods of Computing Assessor Reliability Across 10 Simulated Samples for One Exercise .....	17
Table 3. Average Assessor Reliability Across Three Administration Cycles (2002-2005) for EA/Math and MC Gen .....	18
Table 4. Estimates of Reliability and Decision Accuracy Across Three Administration Cycles .	19
Table 5. Average Exercise Reliability Across Three Administration Cycles (2002-2005) for EA/Math and MC Gen .....	21
Table 6. The Impact of Average Decision Accuracy Across Three Administration Cycles (2002- 2005) for EA/Math and MC Gen.....	22
Table 7. Average White-African American Subgroup Differences Across Three Administration Cycles (2002-2005) for EA/Math and MC Gen .....	23

## List of Figures

Figure 1. Content standards development process. ....	3
Figure 2. Double scoring reliability study design.....	14
Figure 3. Modified double scoring reliability study design. ....	14

# REVIEW OF THE PSYCHOMETRIC QUALITY OF THE NATIONAL BOARD ASSESSMENTS

## Purpose

The National Board of Professional Teaching Standards (NBPTS) awards certificates in 24 different areas of teaching. Thousands of teachers nationwide participate in the assessment process every year. Board certification, for those who pass, is a prestigious recognition of teaching skills and can make those teachers eligible for higher pay or better jobs. Clearly, the NBPTS certification process is important; it affects the lives of thousands of teachers every year and likely has ancillary effects on students.

The importance of teacher assessment and its outcomes led the National Research Council (NRC) to establish the committee on the evaluation of teacher certification by the NBPTS. That committee contracted with the Human Resources Research Organization (HumRRO) to review the psychometric quality of NBPTS assessments. The committee asked us to evaluate the psychometric quality, generally, across all certificates and to report specifically for two certificates selected as examples—Early Adolescent/Mathematics (EA/Math) and Middle Childhood Generalist (MC Gen).

## Approach

Our review had two primary components. One component involved reviewing documents describing the process of assessment development to learn more about the content validation of the assessments. The other component involved summarizing and evaluating data on the technical quality of assessments across three administration cycles. Both components of the reviews involved examining a number of historical technical documents and publications to gain understanding of the evolution of the assessment process. The key documents we reviewed included the following:

- § Annual “Assessments Analysis Report” for three administration cycles (i.e., 2002-2003, 2003-2004, 2004-2005).
- § A March 2007 technical report describing the assessment process.
- § Information provided by NBPTS staff in response to specific questions.
- § Presentations made at a committee meeting.
- § Reports prepared by NBPTS’s Technical Advisory Group (TAG).

A number of professional associations concerned with measurement have developed standards to guide assessment programs including (1) *The Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999); (2) *Principles for the Validation and Use of Personnel Selection Procedures* (SIOP, 2003); (3) *National Commission for Certifying Agencies* (NCCA), and *American National Standards Institute* (ANSI). The standards articulated in these various documents are tailored to different contexts, but they share a number of common features. With regard to credentialing assessments, they lay out standards to guide the process of identifying the competencies to be assessed; developing the assessment and exercises; field testing exercises;

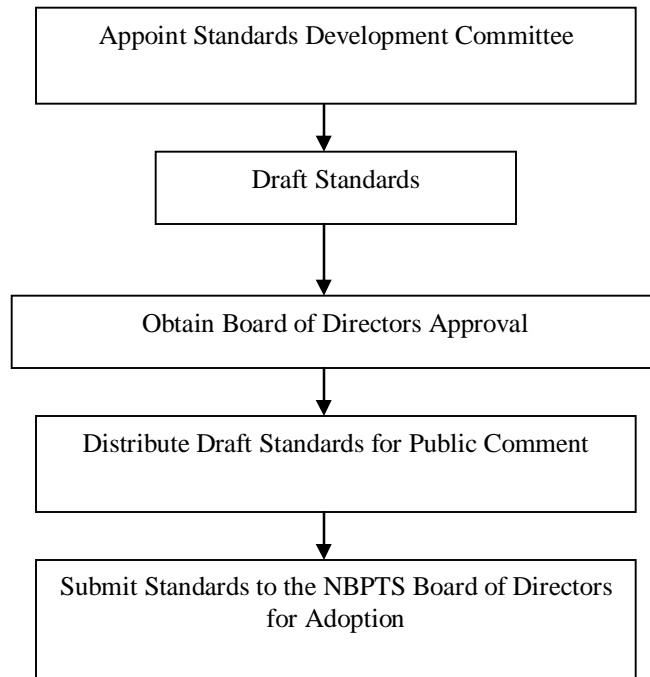
administering and scoring the exercises; setting the passing standard; and evaluating fairness, the reliability of the scores, and the validity of interpretations based on the assessment results. We reviewed the technical documentation on the NBPTS assessments with these criteria in mind.

### **Development, Administration and Scoring of NBPTS Assessments**

This section summarizes the method NBPTS uses to develop, administer and score assessments. The primary source document for this section was a draft technical report (NBPTS, March, 2007). The process begins with the development of content standards for the certificate. Over the years, NBPTS has defined Five Core Propositions for teaching standards regardless of the certification field. The content standards describe how those Five Core Propositions are manifested in each particular certification field. Once the process of developing the standards is in place, NBPTS develops and pilot tests portfolio and assessment center exercise for the certificate. Finally, the exercises are administered to candidates and are scored by trained assessors.

#### ***Development of Content Standards***

The content standards provide the foundation for NBPTS assessments. As such, the standards must accurately reflect the Five Core Propositions and define accomplished teaching in each certification field. The development of the standards is a process that takes 12 to 18 months, sometimes even longer. As depicted in Figure 1, it begins with the appointment of a standards committee by the NBPTS Board of directors. Once selected, the committee meets on several occasions over the course of several months to draft the standards. After the drafted standards have been approved by the NBPTS Board of Directors, they are distributed for public comment and revised. In turn, the revised standards are submitted to the NBPTS Board of Directors for adoption.



***Figure 1. Content standards development process.***

### ***Standards Development Committees***

The effective composition of the standards committees is critical to ensuring that the resulting standards will represent important aspects of teaching in each field. According to the *Standards Development Handbook* (2006), NBPTS posts requests for nominations on the NBPTS Web site, circulates the requests at conferences and meetings, and solicits nominations directly for committee members from disciplinary and other education organizations, state curriculum specialists and chief state school officers, educational leaders, NBCTs, and the NBPTS board of directors. Committee members are selected on the basis of their qualifications and to, on balance, represent diversity on factors such as teaching contexts, ethnicity, gender, and geographic region.

Standards committees are generally composed of 8-10 members who are appointed by the NBPTS board for three years, subject to renewal. Committee members are teachers, teacher educators, scholars, or specialists in the relevant field.

Standards committees interact with other associations, collaborate with standards committees in related fields, and confer with the general contractor on a regular basis. They also confer with other professionals in the field and the general public on the appropriateness of the standards and provide advice on the implementation of the certification process.

### ***Drafting the Standards***

During its initial meeting, the standards committee learns about NBPTS, the Core Propositions, the standards development process, and the structure of a standards development report. Members also participate in discuss several questions about their field (e.g., What are the major issues in your field? What are some individual examples of accomplished practice in your field?). An initial standards document is drafted by a writer who translates the committee's consensus into draft standards. The draft standards are circulated between meetings and are the focus of the next meeting. The process of meeting, redrafting, and re-circulating standards is repeated until the draft standards are ready for submission to the Board.

### ***Getting Input on the Standards***

When the draft standards have been approved by the board, they are released for public comment. The standards are posted on the NBPTS Web site and are distributed directly to educators and leaders of disciplinary and specialty organizations. The public comment period lasts about 30 days. The comments are summarized and circulated to the committee. The committee meets again to review the comments and revise the standards document.

### ***Publishing and Updating the Standards***

The standards are submitted to the board, and after adoption are published. They are available for download at the NBPTS Web Site ([www.nbpts.org](http://www.nbpts.org)). NBPTS views the standards as living documents (NBPTS, 2006). Toward that end they are periodically reviewed to determine the need for revision.

### ***Development of Assessment Exercises***

NPBTS (2007) has two stated main objectives for assessment development: (a) to provide a basis for making valid and reliable judgments about teachers' accomplished practice and (b) to provide opportunities for candidates to develop professionally through participation in the process. Toward those ends, NBPTS has developed a number of principles that guide the development of assessments; they are listed in Table 1.

### ***Table 1. General Assessment Development Guidelines***

---

- A certificate's Standards provide the basis for the development and scoring of all exercises and are the only lens through which a candidate's teaching is assessed.
  - No exercise attempts to elicit evidence pertaining to all Standards.
  - Taken together the exercises collectively assess the domain of accomplished teaching as defined by the Standards.
  - All instructions given to teachers have been written to make clear exactly what is expected from them in terms of what to include in a response to an exercise and how it will be scored. Teachers should not have to guess what information is wanted.
  - Every aspect of teaching cannot be assessed via this form of assessment. There are some aspects of accomplished practice that cannot be documented well within this framework (e.g., personality or dispositional qualities).
  - The assessments focus on the teacher's practice. No specific contextualized factors, such as student ability or wealth of the school district, are valued over others. However, all teaching is interpreted in light of the context in which the teacher works.
  - The assessments are designed to be sensitive to the multiple ways that the accomplished practice manifests itself.
  - A compensatory scoring model is used so that a poor performance in one exercise can be compensated by better performance in another. There are no minimal performance requirements for any single exercise in order to receive National Board Certification.
- 

*Note.* From "National Board for Professional Teaching Standards Technical Report Draft" by the National Board for Professional Teaching Standards, 2007.

### ***Teacher Assessment Development Teams***

Although specific development steps have varied across the years, involvement of teachers in the development, design, and piloting of exercises has always been a key component of NBPT practice. The current practice, for the development of the Next Generation<sup>1</sup> assessments, is to form Teacher Assessment Development teams (TADT) composed of two subgroups: a National Teacher Assessment Development Team (NTADT) and a Local Teacher Assessment Development Team (LTADT) for each certificate. NBPTS recruits participants from randomly selected states and school districts from across the country and solicits nominations from professional organizations, other teachers who had been involved in previous assessment development, and self-nominations. Nominees submit biographical information with a letter of interest, district recommendation, and a resume. In selecting team members, NBPTS considers grade level, subject areas, race/ethnicity, gender, and professional qualifications. Additionally, NBPTS requires that all members be practicing teachers in the subject area and developmental level appropriate for the certificate. For each certificate, NBPTS selects four NTADT members, four LTADT members, and a Teacher-in-Residence.

Members of the TADT and the lead developer from the NBPTS contractor draft portfolio and assessment center exercises and scoring. The process unfolds as follows:

---

<sup>1</sup> In 2002, NBPTS standardized the structure of assessments across all certificates. Assessments after this time were termed "Next Generation."

Test developers present a standards review, refine draft specifications of the assessment based on the Standards, then one draft exercise is reviewed and discussed. Following this discussion, the second and third exercises are drafted based on initial specifications. As part of the development process, each member of the TADT tries out initial draft versions of the portfolio and assessment center exercises and recommends modifications to the exercises. After the modifications are incorporated into the exercises, the members try out the revised draft versions, and, if necessary, recommend additional modifications.

Typically, the TADT meets once each month for a period of 10 months to construct exercises and rubrics. All memos, letters, notes, and biographical information forms and resumes are stored electronically or in binders.

### ***Portfolios***

The portfolios are designed to show how the candidates apply knowledge and theory in real-time, real-life settings. The classroom-based entries are specifically designed for each certificate. They require the candidate to present direct evidence of the teacher's in-class practice and to provide commentary on this evidence. For the documented accomplishments exercise, candidates describe their accomplishments that impact student learning.

In addition to the guidelines in Table 1, three general rules guide the development of portfolios (NBPTS, 2007, p.15-16):

- Classroom-based portfolio entries are designed to elicit evidence pertaining to multiple Standards.
- The responses to all the portfolio entries represent a sample of a teacher's practices. The responses do not sample typical performance, but rather best performance. There is no means to determine whether the practices observed in the portfolio are typical.
- The portfolio assessment is grounded in practice. Teachers are asked to carry out activities that could reasonably be expected to be performed by all teachers.

### ***Assessment Center Exercises***

The assessment center consists of six 30-minute exercises. Assessment center exercises are designed to elicit evidence pertaining to one or more Standards, primarily relating Standards concerned with subject matter content. For generalist certificates, the standards usually stress breadth of knowledge while the standards for specialized certificates place greater emphasis on depth of knowledge in the content area. As needed, the development teams have gathered reference materials for specific knowledge areas for other disciplines.

The current process of developing assessment center exercises begins with the identification of the critical areas of content knowledge described in the Standards to be assessed across the six exercises. The TADT develops a shell or blueprint for the exercise. The shell specifies fixed and variable elements. The fixed portions usually include the instructions, methods, portion of the content domain being samples and the response evaluation criteria. The shells must be specific enough to facilitate the creation of variants on the exercise. According to NBPTS (2007),

Assessment shells are constructed by an analytical methodology that must be followed *every single time* in the same sequence. First, the decision is made on what is to be measured. Then, an analysis of what the credible and appropriate evidence for that particular quality, ability, skill, or behavior would be is conducted. Then an exercise is drafted to capture that evidence (and only that evidence). Finally, the evidence captured with the exercise is reviewed during a tryout phase. The comparative analysis of what was intended to be measured and what was actually obtained is part of the validity evidence that is captured as the assessment is designed (p. 23).

To enhance the security of the assessment center exercises, the contracting team develops variants of exercises by altering the variable portions of the shell within specified constraints (e.g., word length of passages, familiarity of material). Different variants are used across years for all certificates. Multiple variants are used within a single administration year for large volume certificates.

In 1998, the Educational Testing Service (ETS) assessed the equivalence of assessment variants by comparing the scores on variants of four exercises for two certificates, Early Childhood Generalist and Middle Childhood Generalist (NBPTS, 2007, Appendix 11). There were moderate (one-third to one-half *SD*) differences in mean scores for variants of the same question for two of the four Early Childhood Generalist assessments.

### ***Pilot Testing and Formative Scoring of Assessments***

The final versions of the exercises are pilot tested on a sample of teachers who did not participate in assessment development. The objectives of the pilot test are to (a) determine whether instructions are clear and whether the exercises are in need of modification and (b) to estimate time to complete the exercise. Toward that end, the TADT reviews feedback forms from the pilot test and conducts formative scoring of the pilot test responses. Formative scoring is the process NBPTS uses to (a) illuminate problems in the prompts or scoring materials and (b) create final scoring rubrics and other features of the scoring system. As they review responses, the TADT members are asked to pay particular attention to connections between the prompts, the evidence, and the rubric and to identify areas where changes need to be made to the prompts or rubrics in light of the presented evidence. This process identifies prompts that do not elicit the desired evidence and allows refinement of the rubrics to reflect presented evidence.

The TADT categorize responses from the pilot test as representing either accomplished or not accomplished performance. Working in small teams, each handling three exercises, they identify an accomplished response. Next, each team member reviews two responses (1 per variation) for each of the three exercises that they have not previously seen.

The exercises are also submitted to internal and external review to ensure that the exercises to verify the appropriateness and relevance. The NBPTS Board of Directors reviews and approves the final operational version of each full assessment prior to its release.

The first operational year in which an assessment is offered is referred to as the prototype year. During that year, the assessment is double-scored to allow for a better measure of the effectiveness of the scoring system.

### *Administration of Assessments*

#### *Portfolios*

Candidates complete portfolios on their own in accordance with portfolio instructions that are posted at the NBPTS Web site. Over the years, NBPTS has developed materials to help candidates develop their portfolio entries and to make sure that candidates understand what is being asked of them. All certificates use the same presentation template. Candidates are allowed to present student work or video recordings in Spanish for all certificate areas except English Language Arts provided that they also submit an English-language translation of the material. Candidates must submit their portfolios at the end of March, the year following application.

#### *Assessment Center Exercises*

NBPTS administers assessment center exercises at over 400 computer-based testing centers across the United States.<sup>2</sup> Each testing center meets the federal and state requirements for testing services under the Americans with Disabilities Act (ADA) and has an established secure testing network. To accommodate teachers with varied computer experience, the testing session begins with an on-screen tutorial of computer basics, and the system uses a word-processing system that has limited functionality. Any handwritten materials such as graphs or diagrams that candidate's wish to include are

---

<sup>2</sup> In 2003, NBPTS eliminated the option of completing the assessment in handwritten form to minimize costs and potential errors in scoring and processing responses. To ensure that the computer-response requirement would not adversely affect minority candidates, NBPTS compared performance data on two large-volume certificates. For both African American and White candidates, results showed that there were no significant differences in performance on the assessment center exercises for those candidates who chose to handwrite and those who chose to type their responses.

scanned in and attached to the candidate's testing record.<sup>3</sup> Scientific calculators are provided on the computer, and no other calculators or electronic devices are permitted in the testing session.

Initial check-in at the testing center takes about 15 minutes. After the computer basics tutorial, the first exercise is presented. Each exercise follows the same presentation structure. The exercise title and a reminder to scroll to the end of the screen to read all of the information appear. The system presents the instructions and description of the exercise and number of prompts. The criteria, that is, the bullet points outlining the highest level of performance in the scoring rubric, are presented. Finally, the system presents instructions for typing a response, stimulus material (if applicable) and the prompts for each exercise.

Candidates may take a 15-minute break between Exercise 3 and Exercise 4. After completing all the exercises, candidates are asked to respond to a brief survey about their assessment center experience and receive information on how to contact NBPTS and ETS if needed.

Candidates taking the music assessment or assessments for world languages other than English take exercises that require technology that cannot be administered at the testing centers. At locations of their choosing, music candidates listen to stimuli presented on CD and provide handwritten responses to the exercises. Similarly, candidates for world languages other than English have one exercise that must be administered on cassette recorders.

### ***Benchmarking***

Benchmarking is the process NBPTS uses to develop materials that will, in turn, be used by trained scorers to score the assessments. The primary objective is to select examples cases that will anchor the scoring scale for assessors and facilitates assessor training. Trained assessors and other experts participate in 7-8 day benchmarking sessions for portfolios and 3-9 day sessions for assessment center exercises. The participants receive training, generally about NBPTS assessment, and then specific to the portfolios or assessments. After selecting benchmarks (responses that anchor score values), the participants spend several days reviewing cases from the response pool for the exercise. In doing so, they identify scoring issues specific to the exercises and develop training that must be addressed with assessors. Participants look, in particular, for examples that highlight scoring issues. Benchmarks are validated by asking one or more reviewers (who were not benchmark participants) to assign scores to the benchmarks. If differences between the benchmark and reviewer rating are not resolved, the response is not used as a benchmark.

---

<sup>3</sup> Candidates were allowed to respond in handwritten form or via computer through the 2002-2003 administration cycle. Through 2000, all handwritten responses were transcribed, verbatim, and the transcription was checked twice before scoring. In 2001 responses were photocopied, and in 2002 handwritten responses were scanned.

### ***Scoring of Assessments***

Portfolio assessors receive three days of training, and assessment center assessors receive two days of training. Training includes (a) review of the standards and scoring rubrics, (b) practice scoring several training cases, (c) bias training to sensitize assessors to situations when their personal preferences could affect the assessment, and (d) training in the use of a note-taking guide.

All candidate submissions for certificate areas are scored by two assessors during the first administration year for that certificate. After the first administration year for a certificate, “Modified Double Scoring” is used. Twenty-five percent of all cases submitted for a portfolio entry or assessment center exercise are double scored. When a submission is double-scored and the two scores differ by 1.25 points or less, the final exercise score is the average of the two scores. If the two scores differ more than 1.25 points, the lead trainer assesses the response and assigns a score. The final score is computed by either (a) weighting the lead trainer’s score twice as much (if the three scores are less than 2 points apart) or (b) eliminating one rater’s score (if it differs from the others by 2 points or more).

### ***Small Sample Scoring***

When certificates have fewer than 80 responses, the pool of responses is not sufficient for the development of benchmarks and training samples to anchor the scoring system. In these cases, NBPTS uses a consensus scoring system. Three assessors (a) independently read and score a case, (b) meet to discuss the case and the evidence identified, and (c) reach an agreed-upon score. Consensus scoring is usually conducted during benchmarking sessions when trainers and validators are both present to contribute to and monitor the process. Following the scoring of the consensus cases, assessors move to independent scoring, like they do when large samples are scored.

### ***Content Validity Studies***

The content validity of an assessment rests on the extent to which the assessments contain important content to be measured. In the case of the NBPTS assessments, there are two inferences to be validated. One is that the standards for the certificate are relevant to accomplished practice in the certificate area. The other, in turn, is that the assessments are related to the standards. In 1995, the TAG conducted a content validity study of the Early Adolescent Generalist assessment (Loyd, 1995). While the actual content validity evidence in the study is quite dated, because the assessments have changed a great deal since then, it provided a useful methodology for collecting content validity evidence for an assessment. It involved assembling a panel of highly accomplished teachers who had no previous experience with NBPTS assessments. Participants were first asked to assess the extent to which the standards for the certificate reflected critical aspects of highly accomplished teaching practice, thus forming a linkage between the standards and

teaching practice. The next set of questions asked panelists to rate the relevance of the assessment exercise to each of the standard, thus providing a linkage between the assessments and the standards.

The March (2007) technical report describes a content validation study conducted for 21 of the certificate areas in 2002. While the study is not described in detail, it appears to follow the approach used by Loyd (1995). Independent panelists (91% of whom had no prior involvement with NBPTS) were asked to make three types of judgments about the assessments. The first judgment was to rate (on a 5-point scale) the relevance of the standards to accomplished teaching practice. The second set of judgments involved rating rate “how well” the exercises, rubrics, and prompts assess the knowledge, skills, and competencies described in the Standards. The third type of question asked panelists to consider the assessment overall and make judgments about how adequately each of the Standards was assessed. The results were not provided in the March (2007) report; however, the conclusions indicated that the results supported the validity of the assessments.

### **Technical Characteristics: Reliability, Fairness, and Validity**

#### ***Reliability***

Each time NBPTS administers a certificate, a new pool of candidates completes assessments, a new pool of assessors scores the assessments, and some of the assessments may have changed or been replaced. In each case, measurements on the assessment system should be consistent, or repeatable when applied to new candidates, different assessors, in a new year, or even if some exercises are replaced or modified. Reliability refers to the extent to which scores are consistent across different measurement conditions.

No matter how carefully they are developed and scored, assessments are imperfect. Some error is random. Other sources of error can be attributable to different conditions of measurement called facets (i.e., assessors, exercises, occasions). The basics of two theories are important for understanding sources of error and the estimation of reliability—Classical Test Theory (CTT) and Generalizability Theory (G Theory; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). In CTT, a candidate’s observed score on an assessment is equal to his or her true score on the test plus error. G Theory builds on CTT by distinguishing between error arising from different measurement facets (e.g., error arising from assessors, error arising from occasions of measurement).

G Theory requires clear thinking about the intended generalizability of scores across different facets. Specifically, one needs to determine what potential facets of measurement to generalize across. For example, one may want to generalize a measure to new sets of assessors, different measurement occasions, and/or different exercises. If the levels of the facet represent a broader universe or domain, the facet is considered to be a random effect. Effects due to assessors are typically treated as random effects because the results should generalize to a larger population of trained assessors. Effects from

exercises could be considered fixed if the exercises are the only exercises of interest. If exercises represent samples from a larger domain, their effects are random. Decisions about whether effects are fixed or random are important because they affect the estimation of reliability.

### ***Potential Sources of Measurement Error***

NBPTS (2007) describes four potential sources of measurement error in the NBPTS assessments. Two important ones are routinely evaluated by NBPTS—assessor error and exercise error. Another two are not evaluated in NBPTS analyses—school setting and assessment occasions.

#### ***Assessors and Exercises***

Assessor reliability is always a concern when human judgment enters the scoring process. Most assessment programs want to minimize any error that individual assessors contribute to the scoring process. Another source of error NBPTS addresses has to do with the selection of exercises for an assessment. NBPTS materials (2007) note that exercises could be considered as either fixed or random factors. If the exercises are thought of as samples of exercises from a broader domain, it is appropriate to ask how whether candidates would perform similarly on alternate forms of the exercises. If the exercises are the only ones of interest, they would be considered to be fixed element. The latter suggestion might be appropriate for portfolios, but assessment center exercises, which are administered in variant forms should be considered as representing a broader domain.

Through assessor training and other features of the scoring system, NBPTS attempts to minimize potential error such as (a) inferences that assessors might make about the performances or the scoring rubrics and (b) personal biases that assessors bring to the assessment process. Assessor training plays an important role. As mentioned earlier, portfolio assessors receive three days of training, and assessment center assessors receive two days of training. Training includes (a) review of the standards and scoring rubrics, (b) practice scoring several training cases, (c) bias training to sensitize assessors to situations when their personal preferences could affect the assessment, and (d) training in the use of a note-taking guide. Other steps such as conducting read-behinds, re-training, double-scoring and tracking assessor drift also help minimize error in assessments.

In 1996-1997 NBPTS attempted to improve the measurement quality of exams by (a) modifying exercise instructions for candidates (to reduce the amount of guessing that candidates might have to do in preparing responses) and (b) revamping assessor training, scoring materials, and the scoring process. Wolfe and Gitomer (2001) compared the results of Early Childhood/Generalist examinations in 1995-1995 before changes ( $n = 234$ ) to those after implementation of the new procedures in 1996-1997 ( $n = 186$ ). The results suggested that the changes improved the quality of the assessments; inter-assessor

agreement and internal consistency were higher for the 1996-1997 cohort and overall estimates of error variance were lower.

### ***School Setting and Assessment Occasions***

The two potential sources of error described by NBPTS (2007) that are not evaluated are school setting and assessment occasions. Ideally, the assessment should not be a function of where teachers work. According to NBPTS, candidates would need the opportunity to demonstrate accomplished practice in multiple school settings to allow estimation of this source of measurement error.

Another source is assessment occasions. If the scores on the assessment reflect a stable measure of accomplished teaching, it should not matter when the candidate completes the assessment. NBPTS explains that assessing reliability across occasions (a) is not feasible because it would require candidates to complete the entire assessment on two separate occasions and (b) would not be a good way to estimate stability of the assessment because candidates could be expected to learn from the experience of participating in it the first time.

Brennan (2001) has expressed concern that the occasion facet is often overlooked even when it is clear that the organization views its scores as generalizable across occasions. It could be argued that assessment occasions could be assessed, at least in part, with designs that are practical. For example, the National Board of Medical Examiners which uses complex performance assessments for medical licensure testing has taken occasions into account by having assessors score a recorded performance on two separate occasions (Clauser, Clyman, & Swanson, 1999). This design is not a complete replication of the assessment process but does get at the question as to whether the same performance may have been evaluated differently on separate occasions.

### ***Estimating Reliability***

NBPTS uses three indices to estimate the reliability of the assessment system (a) an assessor reliability estimate, (b) the adjudication rate, and (c) an exercise reliability estimate.

#### ***Assessor Reliability***

The design of the assessment process dictates what types of measurement errors can be estimated. During the first administration year for a certificate, all candidates' exercises are scored by two assessors. The general study design for double-scored assessments is illustrated in Figure 2. This design is referred to as a *partially nested* design because every candidate is observed on every exercise by two assessors, but different assessors score each exercise.<sup>4</sup> Assessors are nested in exercises, and exercises are crossed with candidates (i.e., [a:e] x c). An additional complication for NBPTS is

---

<sup>4</sup> Specifically, it is design V-A (Cronbach, et al., 1972, p.36).

that, within each exercise, different assessor pairs score different candidates. That is, assessor “a” is not the same person for each candidate. In this kind of design, it is common to refer to the assessor facet as a “quasi-assessor” facet.

		Exercises							
		1		2		...		10	
		a	b	c	d	...	...	s	T
Candidates	1	X	X	X	X	X	X	X	X
	2	X	X	X	X	X	X	X	X
	3	X	X	X	X	X	X	X	X
	4	X	X	X	X	X	X	X	X
	5	X	X	X	X	X	X	X	X
	...								
	n								

**Figure 2. Double scoring reliability study design.**

As mentioned, after the first year of the assessment, only 25% of the responses are double scored. This type of measurement design sometimes occurs in applied settings (Putka, McCloy, & Le, 2007). In assessment centers, for example, assessors have at times had to recuse from assessing people they already know. Or, when performance ratings are being made it can be difficult to obtain multiple supervisor or peer rating for employees whose jobs are fairly isolated. This design, illustrated in Figure 3, has been called an “ill-structured measurement design” (Putka et al., 2007). There are no well established procedures for estimating reliability for ill-structured measurement designs.

		Exercises							
		1		2		...		10	
		a	b	c	d	...	...	s	t
Candidates	1	X	X	X		X		X	
	2	X	X	X		X		X	
	3	X		X		X		X	X
	4	X		X	X	X		X	
	5	X		X		X	X	X	
	...								
	n								

**Figure 3. Modified double scoring reliability study design.**

To accommodate this situation, NBPTS computes assessor reliability by estimating the (a) total error variance between scorers and (b) the total variance and then entering those values into the CTT formula for reliability (Guilford, 1954, p. 351):

$$\text{Reliability} = 1 - [s_e^2 / s_t^2]$$

where  $s_e^2$  is the total error variance and  $s_t^2$  is the total observed score variance.

Following that general notion, NBPTS (2007) follows six steps to estimate the reliability for a single exercise:

1. Estimate the error variance of the first rating for the group of double-scored candidates—First, use the correlation between the first and second rating in the double-scored group as the measure of reliability. Place this value in the CTT reliability formula shown above along with the variance of first rating in this group and solve for the error term.
2. Estimate the error variance of the second rating for the group of double-scored candidates—Repeat Step 1 replacing the variance of the second rating for the variance of the first rating.
3. Compute the total error variance for the group that was double scored.
4. Estimate error variance in the group that was single-scored—Use the estimate for the first rating of the double-scored group (from Step 1) unless it is larger than the observed variance for the single-scored group. Use the lesser of the two values.
5. Compute the total error variance for the total sample. Calculate the weighted sum of (a) the error variance in the group who was double-scored (computed in Step 3) and (b) the error variance in the single-scored group (computed in Step 4), weighting the error variance of the double-scored group by .25 and the error variance of the single-scored group by .75.
6. Use total error variance from Step 5 as an estimate for  $s_e^2$  in the CTT reliability formula noted above, and divide by total variance to compute reliability. There are a couple of ways that  $s_t^2$  could be computed but NBPTS materials do not describe the method used.

While NBPTS' method of estimating assessor reliability seems reasonable, it is neither state-of-the-art nor status quo. The most common method is to use intra-class correlations (ICCs; McGraw & Wong, 1996; Shrout & Fleiss, 1979) to estimate assessor reliability for assessment center exercises and complex performance assessments. The differences between NBPTS' method and commonly reported methods make it difficult to compare NBPTS reliabilities to those from other complex assessments.

The results of a small-scale simulation conducted for the committee suggest that the NBPTS assessor reliability estimate (a) may be a slight underestimate of the true reliability and (b) is slightly more negatively biased and more variable than other methods of estimating reliability. The simulation used methods developed by Putka et al. (2007). Simulated data were generated for five conditions. In all conditions, the true reliability was set at .70, but the conditions differed in their allocation of variance attributable to assessor main effects (e.g., leniency/severity differences) and variance attributable to the combination of assessor-by-candidate interactions, and residual error. Two samples of 2,000 candidates and 50 raters were generated for each condition. Each simulated candidate had data from two randomly assigned assessors from the set of 50. The second assessor's data was eliminated for 75% of the candidates in each sample to create a group that was single-scored. Restricted maximum likelihood (REML) variance

components (Marcoulides, 1990) were computed using three different random effect models:<sup>5</sup>

- (a) a model in which assessors were treated as nested within candidates;
- (b) a model in which assessors were treated as crossed with candidates, and randomly assigned to be “assessor 1” or “assessor 2”;
- (c) a model in which assessors were treated as crossed with candidates, yet raters were correctly identified (e.g., assessor A, B, C, D)

Based on variance component estimates from the aforementioned models, ICCs were calculated along with the estimated NBPTS assessor reliability. The correlation between the two ratings ( $r_{QR1,QR2}$ ) based on random assignment of assessors to “quasi-assessor 1” and “quasi-assessor 2” columns also appear in Table 2 for reference, and serves as another potential estimator of single-assessor reliability.<sup>6</sup> The single-assessor reliabilities for all methods were, on average across the 10 samples, all .69. The NBPTS estimate was slightly more variable and more negatively biased on average (-1.82%) than the other estimates. Similar results were obtained for the multi-assessor estimate ( $R_{kk}$ ), where  $k$  is the harmonic mean of the number of assessors per candidate in the full sample, namely 1.14 (Winer, 1971).

Given that the NBPTS assessor reliability for a single exercise is expected to be slightly lower than more commonly reported ICCs, it appears that the assessor reliabilities for single exercises within certificates are, for the most part, comparable to those reported in the literature for assessment center assessors. Table 3 shows assessor reliabilities for one certificate that has highly reliable exercises, Early Adolescent/Mathematics (EA/Math), and one certificate that has less reliable exercises, Middle Childhood Generalist (MC Gen). They range from .51 to .62 for MC Gen and from .57 to .94 for EA/Math. In comparison, a meta-analysis of assessment center validities (Arthur, Day, McNelly, & Edens, 2003) reported an average assessor reliability of .86 across six studies.<sup>7</sup> In her book chapter on assessment centers, Tsacoumis (2007) reported assessor reliabilities from two assessment centers each including four job simulations. The average single-assessor reliabilities ranged from .54 to .86, with the majority being more than .70. Reynolds (1999) reported results of role play assessor reliabilities for two managerial assessment center studies. Single-rater reliabilities ranged from .63 to .79 and two-rater reliabilities were between .73 and .88. Reported reliabilities

---

<sup>5</sup> Over the last two decades, advances in variance component estimation methods and computing power (Marcoulides, 1990) have made it possible to compute variance components for some ill-structured designs and, in turn, input those values into appropriate ICC formulations.

<sup>6</sup> It is important to note that different assignments of assessors to quasi-assessor 1 or 2 yields different correlations. The assignment of assessors makes a difference.

<sup>7</sup> The authors did not report whether this was a multi-rater or single-rater reliability.

**Table 2. Comparison of Methods of Computing Assessor Reliability Across 10 Simulated Samples for One Exercise**

Statistic	$r_{QR1,QR2}$	ICC Nested	ICC Crossed (Quasi-Raters)	ICC Crossed (Real Raters)	NBPTS
$M R_{XX}$	.694	.691	.691	.693	.687
$SD R_{XX}$	.023	.021	.021	.021	.028
$M [R_{XX} - R_{XX (True)}]$	-.006	-.009	-.009	-.007	-.013
$SD [R_{XX} - R_{XX (True)}]$	.023	.021	.021	.021	.028
% Bias $R_{XX}$	-.898	-1.248	-1.215	-1.039	-1.824
$M R_{KK}$	.721	.719	.719	.720	.715
$SD R_{KK}$	.022	.020	.019	.020	.026
$M [R_{KK} - R_{KK (True)}]$	.000	-.002	-.002	-.001	-.006
$SD [R_{KK} - R_{KK (True)}]$	.019	.015	.016	.016	.021
% Bias $R_{KK}$	.001	-.328	-.297	-.131	-.880

*Note.* The variance of the “final” score was used as the total variance in the NBPTS formula, where the final score is the average of assessor 1 and assessor 2 scores for double-scored examinees and the score from assessor 1 for the single-scored examinees. Estimates based on the NBPTS formula were  $R_{KK}$  estimates that were adjusted downward to obtain the  $R_{XX}$  estimate using the Spearman-Brown formula.

for performance assessments from educational or credentialing programs have been variable. In a review of the psychometric characteristics of performance assessments, Dunbar, Koretz, & Hoover (1991) reported interrater reliabilities ranging from .33 to .91 across 9 studies. The highest reliabilities were attributable to the use of clearly specified rubrics the lowest reliabilities were found when such rubrics were not used.

When we analyzed data across administration cycles and certificates, we found that the assessor reliabilities are typically higher for assessment center exercises than for portfolios. As shown in Table 3, the MC Gen certificate data are not consistent with the general trends across certificates. That is, the assessor reliabilities for the assessment center exercises are about the same as those for the certificate’s portfolios. EA/Math reliabilities are consistent with the trend across certificates. The median assessment center reliability for EA/Math was .86 compared to the median portfolio reliability of .66.

Previous paragraphs described the reliability of single exercises for certificates. The reliability across all 10 exercises is of particular interest because the overall score is the one used to make decisions about candidates. Once the assessor reliabilities are computed for each exercise, NBPTS applies weights to compute the reliabilities of composite scores across portfolios, across assessment center exercises and across all 10 exercises. As shown in Table 4, the average assessor reliability for the total score across 24 certificates for three administration cycles was .85 with a range of .76 to .93.<sup>8</sup>

<sup>8</sup> Since this estimate is weighted to adjust to proportions of the sample with double-scoring (.25) and single-scoring (.75), it should be considered a multi-rater reliability where the number of raters is on average 1.25.

**Table 3. Average Assessor Reliability Across Three Administration Cycles (2002-2005) for EA/Math and MC Gen**

Exercises	Type	Average R <sub>KK</sub>
<i>Early Adolescence/Mathematics</i>		
Developing and Assessing Mathematical Thinking and Reasoning	Portfolio	.65
Instructional Analysis: Whole Class Mathematical Discourse	Portfolio	.57
Instructional Analysis: Small Group Mathematical Collaboration	Portfolio	.67
Documented Accomplishments: Contributions to Student Learning	Portfolio	.63
<i>Median Portfolios</i>		.66
Algebra and Functions	Assessment	.94
Connections	Assessment	.80
Data Analysis	Assessment	.85
Geometry	Assessment	.86
Number and Operations Sense	Assessment	.94
Technology and Manipulatives	Assessment	.73
<i>Median Assessment Center Exercises</i>		.86
<i>Middle Childhood Generalist</i>		
Writing: Thinking through the Process	Portfolio	.59
Building a Classroom Community through Social Studies	Portfolio	.53
Integrating Mathematics with Science	Portfolio	.54
Documented Accomplishments: Contributions in Student Learning	Portfolio	.58
<i>Median Portfolios</i>		.56
Supporting Reading Skills	Assessment	.53
Analyzing Student Work	Assessment	.54
Knowledge of Science	Assessment	.62
Social Studies	Assessment	.56
Understanding Health	Assessment	.51
Integrating the Arts	Assessment	.59
<i>Median Assessment Center Exercises</i>		.55

### ***Adjudication Rate***

When ratings on double-scored exercises differ by 1.25 points or more (on a scale that ranges from .75 to 4.25), the case is flagged for adjudication. The adjudication rate is thus a simple index of absolute agreement between two assessors. As shown in Table 4 on average, across three administration cycles and 24 certificates, the adjudication rate was 3.3% (for the 25% of cases that were double scored). There are no good published data that can be used to assess this rate. It does, however, raise the issue about what this means for the 75% of cases that were not double-scored. The data suggest that, on average, two assessors would have disagreed by 1.25 points or more on 3.3% of those cases.

**Table 4. Estimates of Reliability and Decision Accuracy Across Three Administration Cycles**

Statistic	2002-2003				2003-2004				2004-2005				Grand Mean
	<i>M</i>	<i>SD</i>	Min	Max	<i>M</i>	<i>SD</i>	Min	Max	<i>M</i>	<i>SD</i>	Min	Max	
<b>Total Score</b>													
<i>N</i> for each certificate	508	655	28	2,557	481	512	54	1,967	480	516	57	1,954	490
<i>M</i>	264	10	244	281	261	9	239	276	260	7	244	274	262
<i>SD</i>	40	4	34	49	39	4	32	48	40	4	33	48	40
Reliability (Exercise formula)	.68	.06	.56	.76	.69	.05	.62	.78	.70	.05	.63	.80	.69
Reliability (Assessor formula)	.84	.04	.76	.91	.86	.03	.79	.91	.86	.04	.78	.93	.85
<b>Percent of Exercise Scores Adjudicated</b>	3.7%	1.0%	2.0%	5.7%	3.4%	1.0%	1.8%	6.2%	2.9%	0.9%	1.6%	5.2%	3.3%
<b>Probability of False Negative Decisions</b>													
Reliability (Exercise formula)	.09	.02	.05	.14	.09	.03	.02	.13	.09	.02	.03	.12	.09
Reliability (Assessor formula)	.07	.02	.04	.10	.06	.02	.01	.08	.06	.01	.04	.09	.06
<b>Probability of False Positive Decisions</b>													
Reliability (Exercise formula)	.10	.02	.06	.14	.10	.02	.07	.18	.09	.01	.07	.12	.10
Reliability (Assessor formula)	.06	.02	.04	.10	.06	.01	.03	.09	.06	.01	.03	.08	.06

*Note.* Reliabilities computed with the “exercise” formula are internal consistency estimates similar to Coefficient Alpha (Jaeger, 1998) and are likely to be conservative. The assessor reliability estimates represent an upper bound on the reliability. “NA” means that the reliability was not available in NBPTS reports. Twenty-five percent of the exercises are scored by two assessors. Exercise scores are adjudicated if assessors disagree by 1.25 points or more on a single exercise. False negative decisions occur when a candidate who should be certified is denied certification. False positive decisions occur when a candidate who should not be certified receives certification.

### ***Exercise Reliability***

NBPTS estimates exercise reliability using an internal consistency estimate developed by Cronbach and reported in Jaeger (1998) and NBPTS (2007). Jaeger describes this coefficient as a form of Coefficient Alpha. NBPTS documents refer to it as an “alternate forms” estimate of reliability because it treats each exercise as a replicate from the broader domain of exercises measuring accomplished teaching. The means of computing exercise reliability reduces the study design to a candidate by exercise matrix by analyzing the candidates’ final scores without regard to raters. This approach estimates how well the scores on nine of exercises predict scores on the tenth. That is, the scores on each exercise are used as measures of a dependent variable, and this dependent variable is regressed on examinees’ scores on all of the other exercises in the assessment. Then, the standard error of estimate (SEE) associated with the regression is used as an estimate of the standard error of measurement (SEM) for the exercise, and in turn, the reliability of the exercise is estimated from the SEM. This process is repeated, treating each of the 10 assessments, in turn, as the dependent variable. A conceptually similar procedure is used to estimate the reliability of the weighted total score across assessment exercises.

The reliabilities for the individual exercises are, in Classical Test Theory terms, reliabilities for a single item on a test. They are consequently low. As shown in Table 4, the average exercise reliability for the total score across 24 certificates for three administration cycles was .69. Compared to alternate forms estimates for multiple-choice cognitive ability measures, .69 is low. However, because internal consistency estimates are tied to the number of items on the test; the internal consistency of a 10-item multiple choice cognitive ability test would also be low. For example, the alternate forms reliability estimate for the Armed Services Vocational Aptitude Test Battery 35-item Word Knowledge subtest is .89 (Palmer, Hartke, Ree, Welsh, & Valentine, 1988). If the Word Knowledge subtest had only 10 items, its estimated reliability would be .61. The average estimate of .69 for NBPTS assessments is, therefore, within the range of what could be expected for a 10-item test.

As with the assessor reliabilities, exercise reliabilities are typically lower for the portfolios than for the assessment center exercises. Again, the MC Gen certificate data do not follow this general trend; the exercise reliabilities for its assessment center exercises are lower than those for its portfolios as shown in Table 5.

### ***Estimating Decision Accuracy***

NBPTS reports the probabilities of erroneous pass/fail decisions for each administration of each certificate. The computation of these estimates is multi-step and “computationally intensive” (Jaeger, 1998 p. 198). It takes into account the reliability of the assessment, the distribution of overall scores, the minimum and maximum possible score, and the performance standard (or cut score) on the assessment. False negative decisions occur when a candidate who should be certified (i.e., has a true score at or above the cut score) is denied certification. False positive decisions occur when a candidate who should not be certified receives certification.

**Table 5. Average Exercise Reliability Across Three Administration Cycles (2002-2005) for EA/Math and MC Gen**

Exercises	Type	Average R <sub>xx</sub>
<i>Early Adolescence/Mathematics</i>		
Developing and Assessing Mathematical Thinking and Reasoning	Portfolio	.21
Instructional Analysis: Whole Class Mathematical Discourse	Portfolio	.14
Instructional Analysis: Small Group Mathematical Collaboration	Portfolio	.20
Documented Accomplishments: Contributions to Student Learning	Portfolio	.17
<i>Median Portfolios</i>		.19
Algebra and Functions	Assessment	.48
Connections	Assessment	.27
Data Analysis	Assessment	.23
Geometry	Assessment	.34
Number and Operations Sense	Assessment	.37
Technology and Manipulatives	Assessment	.27
<i>Median Assessment Center Exercises</i>		.31
<i>Middle Childhood Generalist</i>		
Writing: Thinking through the Process	Portfolio	.21
Building a Classroom Community through Social Studies	Portfolio	.19
Integrating Mathematics with Science	Portfolio	.21
Documented Accomplishments: Contributions in Student Learning	Portfolio	.19
<i>Median Portfolios</i>		.20
Supporting Reading Skills	Assessment	.12
Analyzing Student Work	Assessment	.12
Knowledge of Science	Assessment	.07
Social Studies	Assessment	.09
Understanding Health	Assessment	.14
Integrating the Arts	Assessment	.14
<i>Median Assessment Center Exercises</i>		.12

To get a rough idea of the effect of misclassifications for the NBPTS system overall, these probabilities can be applied to the examinee volume data. Across the three administration cycles, 35,359 candidates completed NBPTS assessments, 13,218 of whom were certified and 22,041 were not certified. On average across administration cycles and certificates, the false negative rates were .06 (based on assessor reliability) and .09 (based on exercise reliability), meaning that 6%-9% of those not certified (22,041) should have been certified (i.e., 1,322 to 1,984). Based on the false positive rates, .06 to .10 of those certified should have been denied. Of the 13,218 who were certified, .06 to .10 should have been denied (i.e., 793 to 1,322). While the rates of misclassification are similar for false-positives and false-negatives, the false negative rate has a greater impact because more candidates fail than pass. Table 6 provides this same type of analysis for the EA/Math and MC Gen certificates.

**Table 6. The Impact of Average Decision Accuracy Across Three Administration Cycles (2002-2005) for EA/Math and MC Gen**

	False Negative Decisions			False Positive Decisions		
	Probability	Number Failing	Decision Errors	Probability	Number Passing	Decision Errors
<i>MC Generalist</i>						
Exercise						
Reliability	.11	4,076	448	.10	2,211	221
Assessor						
Reliability	.08	4,076	326	.07	2,211	155
<i>EA/Math</i>						
Exercise						
Reliability	.07	1,000	67	.09	462	40
Assessor						
Reliability	.04	1,000	43	.05	462	22

### *Fairness*

#### *Subgroup Differences and Disparate (Adverse) Impact*

Two statistical indices are indicators of the extent of subgroup differences in a testing program—the effect size and the disparate impact ratio. The effect size (*d*) is the standardized difference between two groups’ mean scores.<sup>9</sup> It ranges from 0 to 1.0. Large differences between the scores of Whites and African Americans are often observed on tests of cognitive ability or scholastic achievement.

Generally, on the NBPTS assessments, females receive higher scores than males on all types of assessments, but the male-female difference on the assessment center exercises is quite small. Whites receive higher exercise scores than other racial/ethnic subgroups do, and effect sizes for the portfolios are smaller than those for the assessment center exercises. The average White-African effect size across 3 administration cycles and all 24 certificates was .53 for the portfolios and .70 for the assessment center exercises.

Table 7 shows the White-African American effect sizes on individual exercises for the two example certificates. The EA/Math exercise effect sizes follow the general pattern that we observed across all certificates, where the effect sizes for the assessment center exercises (i.e., median = .73) are notably higher than those for the portfolios (i.e., median = .40). This trend also appears for the MC Gen exercises, but the magnitude of the effect size difference is not as large.

<sup>9</sup> (Subgroup 1 Mean – Subgroup 2 Mean)/Pooled Standard Deviation.

**Table 7. Average White-African American Subgroup Differences Across Three Administration Cycles (2002-2005) for EA/Math and MC Gen**

Exercises	Type	Average <i>d</i>
<i>Early Adolescence/Mathematics</i>		
Developing and Assessing Mathematical Thinking and Reasoning	Portfolio	.39
Instructional Analysis: Whole Class Mathematical Discourse	Portfolio	.35
Instructional Analysis: Small Group Mathematical Collaboration	Portfolio	.44
Documented Accomplishments: Contributions to Student Learning	Portfolio	.40
<i>Median Portfolios</i>		.40
Algebra and Functions	Assessment	.75
Connections	Assessment	.54
Data Analysis	Assessment	.70
Geometry	Assessment	.78
Number and Operations Sense	Assessment	.93
Technology and Manipulatives	Assessment	.67
<i>Median Assessment Center Exercises</i>		.73
<i>Middle Childhood Generalist</i>		
Writing: Thinking through the Process	Portfolio	.50
Building a Classroom Community through Social Studies	Portfolio	.46
Integrating Mathematics with Science	Portfolio	.55
Documented Accomplishments: Contributions in Student Learning	Portfolio	.51
<i>Median Portfolios</i>		.51
Supporting Reading Skills	Assessment	.63
Analyzing Student Work	Assessment	.62
Knowledge of Science	Assessment	.61
Social Studies	Assessment	.60
Understanding Health	Assessment	.61
Integrating the Arts	Assessment	.62
<i>Median Assessment Center Exercises</i>		.62

The disparate impact ratio takes into account the passing rate. It compares the percentages of different subgroups who achieved a passing score (i.e., minority subgroup percent passing/majority subgroup percent passing). The legally recognized criterion for disparate impact is referred to as the 4/5<sup>th</sup> rule. If the disparate impact ratio is less than .80, meaning the minority passing rate is less than 4/5<sup>th</sup> of the majority passing rate, the assessment has disparate impact. Across the three administration cycles that we analyzed, the average passing rate was 38% across all certificates. Passing rates for subgroups were 41% for Whites, 12% for African Americans, and 31% for Hispanics. On average across certificates, there is disparate impact for both African Americans and Hispanics, but the disparate impact is much larger for African Americans.

Both of the selected example certificates result in disparate impact for African Americans and usually for Hispanics as well. For the MC Gen certificate, the average overall pass rate was 35% across 3 administration cycles. The African American and Hispanic pass rate were 12% and 21%, respectively, compared to the White pass rate of 38%. For EA/Math, the average overall pass rate was 32%. The White pass rate was 32% compared to 9% for African Americans and 26% for Hispanics. The difference between pass rates for Whites and Hispanics on the EA/Math was relatively small and did not meet the definition of disparate impact.

### ***NBPTS Research on Disparate (Adverse) Impact***

NBPTS has been concerned about disparate impact since the early days of the testing program and has conducted several studies investigating it. The Board members, particularly Lloyd Bond (1998a, 1998b), spearheaded most of this research. He has found that there is no simple explanation for the White-African American race difference. He found that there do not appear to be important differences between the number of advanced degrees and years of teaching experience of White and African American candidates. To investigate the possibility that disparate impact resulted in part from differing levels of collegial, administrative, and technical support, the Board conducted in-depth phone interviews of candidates. In the end, the analyses suggested that the level and quality of support were not major factors in the disparate impact observed.

The Board conducted analyses to assess the idea that disparate impact might be a function of assessor judgments and biases (Bond, 1998a). Initially, they located a small number of cases where African-American and White assessors evaluated the performances of the same candidates of different races. Their analyses revealed that African-American assessors tended to be slightly more lenient overall, but no interaction between assessor race and candidate race was noted. That is, African American candidates who were scored low by white assessors were scored low by African-American assessors as well. Since the initial, small-sample study, the Board has continued to analyze these data where sample sizes permit. Results of the later efforts echo that from the early work. Assessor bias does not appear to be the source of disparate impact.

The Board also investigated the idea that an irrelevant variable (e.g., writing ability) is causing the disparate impact (Bond, 1998a). The Board identified an Early Adolescent/Generalist exercise with significant writing demands and others that did not. They conducted ANOVA analyses to assess the effects of race and writing demands. In both analyses, the main effect of race was significant. The main effect of writing demand was also significant. However the race x exercise type interaction was not significant. Writing ability is not a likely explanation for the disparate impact.

Other investigations have focused on instructional styles and NBPTS's vision of accomplished practice (Bond, 1998b). One study investigated the notion that the teaching style that is often more effective for the children that African American teachers teach is one that is not favored on the assessment. Subpanels of a review team "read across" the portfolios and assessment center exercises submitted by candidates in a study sample (in contrast, assessors typically rate only one exercise). The 15 member panel was divided into 5 groups of 3 assessors. Performance materials of all 37 African American candidates in 93-94 and 94-95 for EA/ELA were distributed to the groups. Assessors reviewed all 37 candidates independently and judged whether the candidate's materials contained culturally related markers that might adversely affect their evaluation of the candidate's accomplishment. Twelve of the 37 were deemed accomplished by at least 1 panel member. Five of 37 had been certified. While this study resulted in a few of the candidates who originally failed as being classified as accomplished, it did not reveal consistent differences in instructional styles for African American teachers.

Another study considered varying views of accomplished practice as a source of subgroup differences (Bond, 1998b). Twenty-five African American teachers and former teachers participated in focus group discussions. They were asked to (a) discuss the scope and content of the NBPTS certification standards and to note how the standards differed from their own views about accomplished practice, (b) discuss the portfolio instructions with a view toward possible sources of disparate impact, (c) apply their own weights to the EA/ELA assessment exercises, and (d) evaluate the small-group discussion exercise component for two candidates. The major conclusions that Bond (1998b) drew from the focus groups are listed below.

- Absent powerful incentives, accomplished African American teachers would generally not seek NBPTS certification for fear of risking their excellent reputations.
- Constraints imposed by districts and by students may work against African American teachers (e.g., district content guides that are in conflict with NBPTS views).
- Given that academically advanced students tend to make their teachers look good, those who teach students who are seriously behind, as many African American teachers do are forced to teach lessons that may appear trivial to assessors.
- There was a concern that some principals keep African American teachers out of the loop regarding professional opportunities.

### **Conclusions**

It is important to note that our psychometric review was not as thorough as would ideally be the case. A key factor in this situation is the sheer number of certificates to be reviewed—24. Unfortunately, we were further limited by a lack of technical documentation about how the standards and assessments for each certificate were developed. Indeed, it was much too difficult to obtain basic information about the design and development of the NBPTS assessments that was sufficiently detailed to allow independent evaluation. In early 2007, NBPTS drafted a technical report that seeks to fill some of the information gaps, but this should have been in place long ago for the program to be in compliance with professional testing standards in this regard (e.g., AERA, APA, NCME, 1999; SIOP, 2003). While the number of certificates makes this documentation requirement challenging, it makes the requirement all the more critical. It calls into question how NBPTS and its contractor can share the information they need to effectively run a program that involves so many assessments and the large number of people who work with them.

The primary questions we addressed in this review were two-fold: (a) does the Standards and assessment development process provide evidence of the content validity of the assessments and (b) are the scores from the assessments psychometrically sound. In general, many of the steps in the development of the Standards and the assessments appear to be strong ones. The care given to develop benchmarks and to train assessors is a strength of the program. The shallow description of the development of assessments is a weakness. Also, we must note, that while the process in general appears to be a sound one, we did not receive decipherable materials that would allow us to assess the execution of the process for specific assessments. Details of assessment development and pilot testing should appear in reports.

In contrast, psychometric information about the assessments is routinely produced in very clear assessment reports. NBPTS, as noted earlier in this chapter, does not use typical formulas for computing reliability coefficients and the actual formulas were not available to us until we received the March (2007) report. Regardless, NBPTS assessor reliabilities are in the ballpark of what would be expected for assessment centers and support the idea that NBPTS's benchmarking and training procedures are sound ones. The exercise reliabilities are within expectations for a test with 10 scores. The reliability is lower than that for most credentialing programs that rely on multiple-choice testing. The concern here is decision accuracy. Lower reliability and higher failure rates on the test, lead to a large number of false negatives (i.e., teachers who should have been certified but were not). These issues are not simple ones and are important considerations for NBPTS decision-makers.

### References

- Arthur, W., Day, E. D., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56*, 125-154.
- Bond, L. (1998a). Disparate impact and teacher certification. *Journal of Personnel Evaluation in Education, 12*, (2), 211-220.
- Bond, L. (1998b). Culturally responsive pedagogy and the assessment of accomplished teaching. *Journal of Negro Education, 67*, (3), 242-254.
- Brennan, R. L. (2000). (Mis)conceptions about generalizability theory. *Educational Measurement: Issues and Practice, 19*, 5-10.
- Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38*, 295-318.
- Brennan, R.L. (2001b). *Generalizability theory*. New York: Springer-Verlag.
- Clauser, B. E., Clyman, S. G., & Swanson, D. B. (1999). Components of rater error in a complex performance assessment. *Journal of Educational Measurement, 36*, 29-45.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley & Sons, Inc.
- Dunbar, S. B., Koretz, D., & Hoover, H. D. (1991). Quality control in development and use of performance assessments. *Applied Measurement in Education, 4*, 289-304.
- Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill Book Company, Inc.
- Jaeger, R.M. (1998). Evaluating the psychometric qualities of the National Board for Professional Teaching Standards' Assessments: A methodological accounting. *Journal of Personnel Evaluation in Education, 12* (2) 189-210.
- Loyd, B. (1995). *Content validation of the National Board for Professional Teaching Standards Early Adolescent Generalist Assessment*. Greensboro, NC: Technical Advisory Group.
- Marcoulides, G. A. (1990). An alternative method for estimating variance components in Generalizability theory. *Psychological Reports, 66*, 379-386.
- McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlations coefficients. *Psychological Methods, 1*, 30-46.
- National Board for Professional Teaching Standards (2006, May). *Standards development handbook*. Chicago: Author.
- National Board for Professional Teaching Standards (2007, March). *Technical report: Draft*. Chicago: Author.

- Palmer, P., Hartke, D. D., Ree, M.J., Welsh, J.R., & Valentine, L. D. (1988). *Armed Services Vocational Aptitude Battery (ASVAB): Alternate forms reliability (Forms 8, 9, 10 and 11)* (AFHRL Technical Paper 87-48). Brooks AFB, TX: Air Force Human Resources Laboratory.
- Putka, D. J., McCloy, R. A., & Le, H. (2007, April). *Analyzing ratings from ill-structured measurement designs*. A paper presented at the 22<sup>nd</sup> Annual Society for Industrial and Organizational Psychology Conference in New York City.
- Reynolds, D. J. (1999). *Assessing the assessor: Understanding the reliability of assessor judgment*. A paper presented at the annual meeting of the International Assessment Center Congress.
- Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420-428.
- Tsacoumis, S. (2007). Assessment centers. In D. L. Whetzel and G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management*. Mahwah, NJ: Lawrence Erlbaum Publishers.
- Wolfe, E.W. & Gitomer, D. H. (2001). The influence of changes in assessment design on the psychometric quality of scores. *Applied Measurement in Education*, 14, 91-107. Detailed cost information is provided in the attachment.