ORIGINAL ARTICLE

ABR Psychometric Testing: Analysis of Validity and Effects

Lincoln L. Berland, MD[a], *Nancy W. Berland, PhD*[b], *Matthew W. Berland, PhD*[c]

Abstract

Changes in the certification program of the ABR were first described in some detail in 2008 and have since undergone refinements. Controversies surrounding these changes have included the relevance of test questions, costs, effects of altering the examination scheduling, residency program curriculum changes, and issues related to preparing for the examinations. However, the role of psychometric testing in radiology itself, as the technical foundation for the new ABR Core and Certification Examinations, has undergone less scrutiny. This article examines the validity and consequences of the ABR psychometric testing process, and we conclude that its validity can be challenged and that negative consequences, including adverse effects on allocating human and financial resources and on what is taught and learned in residency programs, should be addressed. The ABR could collaborate with the ACGME, education experts, patients, and public representatives to reform their testing processes, especially by integrating modern evaluation techniques that more authentically simulate radiology practices to better align the examination with its intended purposes.

Key Words: Radiology board certification, ABR, psychometric testing, residency training

J Am Coll Radiol 2018;■:■-■.

INTRODUCTION

The radiology board certification (BC) process had changed little for several decades, until 2007, when multiple publications from ABR representatives began to appear regarding extensive upcoming changes in certification testing. These were followed by additional details, clarifications, and modifications that have been phased in since 2013 [1-11]. Patterning their changes after the systems of other American Board of Medical Specialties (ABMS) member boards, the ABR eliminated the oral component of the examinations for diagnostic radiology, changed the timing of the certifying examinations, and redesigned their tests to enhance standardization and maximize security by using the principles of high-stakes, computer-based, multiple-choice (MC) psychometric testing.

We support the broad agreement that BC has value. However, over the course of implementing these innovations, multiple publications have documented concerns about the testing instruments the ABR now employs [12-22]. The online Appendix to this article represents a systematic analysis challenging that such tests measure what is intended (defined as the "construct"). The remainder of this article evaluates other topics related to psychometric testing, including: (1) conceptual and practical issues with psychometric MC testing, (2) problems encountered with standardized testing in other environments, (3) the effects of disassociating MC testing from face-to-face examinations, (4) effects on residency programs, (5) the ABR's stance on "recalls" (prior board examination items recorded from memory of subjects), (6) additional testing security requirements, and (7) redirection of human and financial resources.

BASIS OF PSYCHOMETRIC TESTING

The descriptions of psychometric testing by the ABR and ABMS imply that psychometric testing is established science because "three professional organizations" have

[a]University of Alabama at Birmingham, Birmingham, Alabama.

[b]Grayson & Associates, Birmingham, Alabama.

[c]Department of Curriculum & Instruction, School of Education, University of Wisconsin–Madison, Madison, Wisconsin.

Corresponding author and reprints: Lincoln L. Berland, MD, Professor Emeritus, University of Alabama at Birmingham, 3421 Brookwood Trace, Birmingham, AL 35223; e-mail: lberland@gmail.com.

Dr Lincoln Berland is a member of the Board of Chancellors of the ACR. However, these papers do not represent positions of the ACR or the authors' institutions, but are solely the opinions of the authors. The authors have no conflicts of interest related to the material discussed in this article.

1546-1440/18/$36.00 ■ https://doi.org/10.1016/j.jacr.2018.02.023

created "the authoritative set of standards . . . of high-stakes testing," which were first published in 1966 and were revised most recently in 2013 [11]. However, accepting this assertion can be challenged by evaluating its effectiveness in medicine and elsewhere and reviewing concepts of what are considered "authentic" tests.

Psychometric Testing Effectiveness

The kindergarten through 12th-grade (K-12) experience with psychometric testing is vast, and the tests are constructed from the same psychometric principles as the ABR Certification Examinations. Although there are obvious differences between public school education and post-medical school professional education, we believe that evaluating the parallels is worthwhile because both environments use psychometric testing as a "high-stakes" method for establishing "promotion" and because they have both been observed to influence teaching methods, including "teaching to the test," as will be discussed further.

Prominent educators who helped to architect the No Child Left Behind (NCLB) law that was enacted in 2002 include Diane Ravitch, PhD, and Linda Darling-Hammond, EdD. Ravitch notes: "What is tested may ultimately be less important than what is untested, such as a student's ability to seek alternative explanations, to raise questions, to pursue knowledge on his own, and to think differently. . . . [MC tests] are not sufficient to measure student learning" [23].

Focusing on radiology, there has been scant research on the effectiveness of ABR testing, either before or after the elimination of the oral component of the Certification Examination [24,25]. The ABR terminated the oral examination for diagnostic radiology, citing the inability to standardize the face-to-face examination [3]. However, in a seminal article on authentic testing, Grant Wiggins [26] notes that multiple judges can demonstrate high interobserver reliability when trained in authentic assessment and when using common criteria. These issues are discussed further in the online Appendix.

What Standardized, Psychometric Tests Measure

Although widely perceived as "time-tested," standardized assessment originated only about 100 years ago as an offshoot of an industrial management movement intended to improve factory production, which partly explains the somewhat mechanistic approach of psychometric testing for assessing intellectual achievement (eg, rewarding speed of recall, testing subjects in the same way at the same time, and depending on "distractors" to spread scores) [26]. A more modern view is expressed by Wiggins, who contrasted "standardized" and "authentic" tests. He notes that "a standardized test of intellectual ability is a contradiction in terms . . . [and is] . . . irrelevant (and even harmful) to genuine intellectual standards." He believes that "a true test of intellectual ability requires the performance of exemplary tasks. . . . Standardized tests are inherently unable to encompass the inevitable idiosyncratic cases for which we ought always to make exceptions to the rule" [26].

So, what do psychometric MC tests really measure? Walter Stroup has studied the results of standardized testing in the K-12 environment, and indicates the following: "Currently an item or test that looks like it is about the specified domain (has face validity) and that produces consistent results (is reliable) is assumed to be valid. However, if the reliability can be shown to be overwhelmingly anchored in something other than the domain (eg, a relatively domain-general test-taking ability) then the assessment should not be considered valid" (Walter Stroup, PhD, Associate Professor and Chair of STEM [Science, Technology, Engineering and Mathematics] Education and Teacher Development at the University of Massachusetts–Dartmouth, personal communication).

Studies have reported that MC tests are remarkably insensitive to differences in instruction: "Items are selected for their ability to be correlated with one another. This 'agreeing with each other' factor starts to be what the tests are about. Calculations showed that individual student test scores changed within a relatively narrow window of about 10 to 15% over many years and that about 3-8% of results are related to instruction, 50-60% to test taking ability, and the remainder are stochastic (random). That is, even the noise is much larger than the signal" [27]. This helps explain why "many students who do well on school tests seem so thoughtless and incompetent in solving real world problems" [26].

Relationship Between Curriculum Content and "Prepping" for the MC Tests

What seems paradoxical regarding NCLB is that "declines on international tests… occurred even as state test scores used for NCLB were driven upward." [28] It was concluded that "the gains in test scores . . . were typically the result of teaching students test-taking skills and strategies, rather than deepening their knowledge" [29]. In radiology, this test preparation is manifested both