The Validity and Timing of the ABR Core Exam

William D. Kerridge, MD, Richard B. Gunderman, MD, PhD,

Department of Radiology, Indiana University, 702 North Barnhill Drive, Room 1053, Indianapolis, IN 46202

Key Words: Validity; timing; board exam; education

The American Board of Radiology's (ABR) new Core Exam is not working, at least not as well as it needs to. Having helped to prepare candidates (RG), studied for and taken the exam (WK), and talked with hundreds of candidates who have taken the exam (RG and WK), we believe that that one aspect of the exam, its validity, can be significantly enhanced.

Just as we expect candidates for board certification and practicing radiologists to measure up to the high standards, so we should subject the board exam to continuous scrutiny, seeking opportunities to rectify errors and enhance the exam's overall quality, with a view to better promoting excellence in radiology practice and the care of patients. Here we focus on two exam parameters: validity and timing.

The Exam

Taken 36 months after the beginning of radiology residency, the Core Exam is administered over 2 days at either the Chicago or Tucson exam center. According to the board, it “tests knowledge and comprehension of anatomy, pathophysiology, all aspects of diagnostic radiology, and physics concepts important for diagnostic radiology.”

Eighteen categories are included: breast, cardiac, gastrointestinal, interventional, musculoskeletal, neuroradiology, nuclear, pediatric, reproductive or endocrinology, thoracic, urinary, vascular, computed

---

This is the author's manuscript of the article published in final edited form as:

Kerridge, W. D., & Gunderman, R. B. (2016). The Validity and Timing of the ABR Core Exam. Academic Radiology, 23(9), 1176–1179. https://doi.org/10.1016/j.acra.2016.05.004

tomography, magnetic resonance, radiography or fluoroscopy, ultrasound, physics, and safety. The exam is administered twice per year, June and November.

Validity

Simply put, the validity of a test is the extent to which it accurately assesses what it is intended to assess. A valid test is one whose results actually mean something, whereas an invalid test is one whose results fail to tell us much—or positively mislead us—about whatever the test is designed to assess.

A famous example of a test with poor validity was the Scholastic Aptitude Test. After decades of research failed to support the hypothesis that the test actually provided useful assessment of scholastic aptitude, in the 1990s the test's owners finally changed its name to SAT (1). The test is not completely useless, but less than 20% of college performance is predicted by SAT scores.

In radiology, a valid board certification test would provide meaningful insights into a candidate's degree of preparation to practice radiology independently, both avoiding errors that could harm patients and making substantial contributions to patient care. A single question captures this reasonably well: would I want this candidate providing a loved one's radiological care?

A radiology board exam with a high degree of validity would enable experts in the field of radiology to answer this question with confidence. On the other hand, an exam with a low degree of validity would provide relatively little insight on this question, both with respect to individual candidates and to the aggregate of all candidates seeking certification.

Assessing the validity of any test involves a number of factors, including its content, the way test-takers enter responses, the internal structure of the test, and the consequences that are associated with test performance. By tweaking or in some cases overhauling a test, it is often possible to enhance its validity.

In this article, we focus on six types of validity, as outlined in the ABR's own “Quality & Safety Domain Specification & Resource Guide” (2): face validity, construct validity, content validity, concurrent

validity, discriminative validity, and predictive validity. In each case, we address the question, how could the performance of the Core Exam be improved?

Face Validity

Face validity concerns the similarity between the test experience and real-life practice (3). To determine whether the Core Exam exhibits a high degree of face validity, radiologists need to ask themselves what their daily work experience is like, and then compare it with the experience of candidates taking the Core Exam.

Although there is a danger in overgeneralization, the experience of most diagnostic radiologists involves the performance and interpretation of imaging exams that are requested to answer more or less specific clinical questions. For example, a chest radiograph is requested to assess for pneumonia in a patient with cough and fever.

One problem with the Core Exam is its multiple-choice format. In our experience, referring health professionals essentially never request radiological consultation in the form of a multiple-choice question. Other than binary choices such as “pneumonia or not” or “fractured or not,” the radiologist rarely confronts a prescribed range of alternatives, one of which is guaranteed to be true.

In this sense, the face validity of the Core Exam, and for that matter any multiple-choice exam, leaves much to be desired. It offers an inauthentic representation of the uncertainty that practicing radiologists confront not only every day but in every case, where the range of possible diagnoses generally cannot be specified in advance.

To put this somewhat differently, the practice of radiology more closely resembles an essay exam than a multiple-choice exam. The radiologist is required, on the fly, to determine which possibilities are relevant and which are irrelevant, and formulate a differential diagnosis and diagnostic plan, without the benefit of a specified range of alternatives.

Construct Validity

Construct validity concerns the ability of the test to discriminate between candidates at different levels of training. An exam with a high degree of construct validity would enable assessors to differentiate broadly between candidates who are just beginning their training and those who are completing it. If neophytes perform as well as veterans, the test offers little insight into the value of training (4).

In one respect, the Core Exam offers a useful assessment of construct validity. Candidates who have completed 3 years of radiology residency would undoubtedly perform significantly better than those just commencing their training. This suggests that the exam is able, at least to some degree, to measure the development of expertise in diagnostic radiology.

What we don't know, however, is how the construct validity of the Core Exam stacks up against other forms the exam could take. For example, does the Core Exam provide a more accurate assessment of candidate experience as compared with the Oral Exam that it helped to replace? Does the multiple-choice format impose inherent limitations on the discriminative capabilities of the exam?

If the Core Exam provides a high-quality assessment of radiological expertise, then radiologists who have 10, 20, and 30 years of experience in a particular field should perform better than candidates with 3 years of experience. Based on what we hear from candidates, we question whether a substantial percentage of the questions truly assess clinically relevant expertise.

Simply put, conversations with hundreds of candidates who took the exam lead us to suspect that a number of the questions on each exam could be fairly characterized as arcane—useful, perhaps, in discriminating between those who have studied excessively and those who have prepared adequately, but not so helpful in determining whether a candidate is prepared for the real world.

Content Validity

This brings us to the concept of content validity, which the ABR defines as the ability of a test to determine whether candidates possess the necessary knowledge and skills. Others may define content validity differently, for example, the extent to which an assessment represents all facets of tasks within the domain being assessed 5 ; 6. The mere fact that someone scores well on a test does not mean that he or she knows what he or she needs to know, or can do what he or she needs to do, unless the content of the test is valid.

To determine content validity, it is vital to understand the knowledge and skills at the core of a discipline. It might be possible to discriminate between radiology residents at various levels of training based on their level of confidence or how well they dress, but this alone does not establish that they are prepared to practice independently.

Because the Core Exam relies exclusively on a multiple-choice format, there are many important aspects of a radiologist that it cannot assess. For example, it cannot tell us how honest, helpful, compassionate, dedicated, or creative a candidate is. Nor can it tell us how well a candidate communicates or how effectively he or she functions as a consultant.

The Core Exam does not provide a comprehensive assessment of a candidate's desirability from the point of view of a potential patient, partner, or employer. An individual's knowledge and skills bases are important, but just because a particular performance parameter lends itself to a multiple-choice testing format does not make it independently more important than others that do not.

In increasing its commitment to a multiple-choice format, the ABR deprioritized a number of important learning objectives in the minds of its candidates. Whether it intended to or not—whether or not it even foresaw such implications—it shifted the sights of candidates preparing for the board exam away from crucial abilities such as communication and collaboration.

To improve the content validity of the board exam process, it would be necessary to add such content domains, which would in turn require a change in the format of the exam, replacing some multiple-choice

components with others that assess candidates' abilities to function effectively in real time as communicators and consultants.

Concurrent Validity

Concurrent validity refers to the correlation between a test's results and other means of assessing candidate performance. Considering the United States Medical Licensing Exam, for example, we have known students who scored highly but had difficulty finding a residency position and others who scored poorly but easily found a residency position.

How could this happen? Simply put, there are many desirable features of a residency candidate that the United States Medical Licensing Exam does not assess, including many mentioned previously. There are a great many features about any person that no standardized exam can elucidate, indicating that the concurrent validity of such exams is always limited.

The same is true of the ABR Core Exam. We have spoken to numerous residency program directors who have been surprised that some of their better residents did not pass the Core Exam, whereas others that they did not regard so highly passed. There is not a 1:1 correspondence between the quality of residents and their Core Exam performance.

In short, the Core Exam leaves much to be desired as an indicator of a candidate's readiness to practice radiology independently. Although we would certainly look at candidates' board certification status before hiring them into our practice, there is a great deal of additional information we would collect before determining whether to allow them to assume responsibility for our patients.

Of course, simply doing away with the Core Exam has its own drawbacks. Using a standardized exam is appealing, because the pressure on program directors to certify the preparedness of their candidates could lead some to provide such an attestation in cases where it was not truly warranted. Standardized tests confer some measure of objectivity and fairness.

Discriminative Validity

This leads to another dimension of validity, discriminative validity. This describes the distinction between results on the test and the results of other means of assessment. If a test offers no discriminative value compared with other forms of assessment, then there is probably no reason to require the test, as it is adding nothing.

Suppose, for example, that it turned out there was no significant difference between the performance of radiology board candidates on a shortened, 1-day Core Exam and the current 2-day exam. If this were the case, then it would be difficult to justify having the candidates go to the extra time and expense of sitting for a 2-day exam.

We believe that almost any multiple-choice exam of reasonable quality would offer significant discriminative ability, compared with other means of assessing radiology residents. After all, a multiple-choice exam is so different from what radiologists do day in and day out that it could hardly fail to differ from clinically based assessment.

But let's face it, some people are just better at taking multiple-choice tests than others, and this ability does not correlate terribly well with how well they actually perform at their jobs. What we do not know is whether the discriminative validity of multiple-choice testing offers benefits sufficient to compensate for its many weaknesses.

Predictive Validity

It is with predictive validity that the rubber really meets the road, because it concerns the degree to which assessors can tell based on exam performance how well candidates will actually perform in the professional roles for which they have been preparing (7). Are candidates who perform well on the Core Exam most likely to perform well as radiologists?

Here it would be helpful if the ABR were conducting more studies of its exam's predictive validity and sharing these widely throughout the field. Unfortunately, it is difficult for organizations other than the ABR to conduct such studies. It jealously guards exam content and exam results are rightly treated as confidential.

As it is, we simply do not have sufficient evidence to determine whether the Core Exam is successfully distinguishing between the competent and the incompetent, or the excellent and the merely adequate. This much we can say: we know people who have failed to pass the exam on their first attempt who we would hire in a minute, and others who have passed that we would never hire.

Timing

With several years of experience under our belts, we can now say that the timing of the exam is suboptimal. Over and over again at the past two meetings of the Association of University Radiologists, we have heard program directors, faculty members, and radiology residents complaining that with the new exam, residents seem less engaged in learning in their fourth year.

As one program director put it, “The fourth-year residents are now ‘checked out.’ They are learning less than they used to in their final year. They don't attend daily didactic conferences the way they used to, and even when they are there, they are not as engaged. This is largely because they don't have the board exam to study for at the end of their fourth year.”

One way to improve the Core Exam would be to move it closer to the end of radiology residency, to keep residents deeply engaged in learning for a longer span of their training. No college or university course administers its final exams three quarters of the way through the semester, and for a good reason. Right or wrong, the intensity of learning is powerfully influenced by the timing of tests.

Conclusion

The Core Exam is administered every year like clockwork, scores are calculated, psychometricians pore over performance on each question, candidates are pleased or crushed when they receive their results, and careers are launched, postponed, or terminated. But we do not know whether the results are truly valid.

Should candidates who do not pass be barred from practicing radiology independently? Is the exam offering sufficient value to warrant the $2500 candidates pay to sit for it and the $625 a smaller number pay to retake portions of it, to say nothing of the large sums many must pay to travel for the exam and the time lost from training and contributing to patient care?

It lies in the best interests of the profession of medicine, the field of radiology, and the patients we serve to ensure that the board exams in general perform as well as possible at what they really need to accomplish. For this to happen with the ABR Core Exam, rigorous scrutiny and vigorous discussion and debate are called for. It is our hope that this article will help to jumpstart these conversations.

References

1. Commission on New Possibilities for the Admissions Testing Program. Beyond prediction. New York City, NY: College Entrance Examination Board, 1990; 9.

2. American Board of Radiology. Quality & safety domain specification & resource guide. Tuscon, AZ, 2016; 34–35.

3. Weiner IB, Craighead WE. The Corsini encyclopedia of psychology. Hoboken, NJ: Wiley, 2010; 637–638.

4. Devitt JH, Kurrek MM, Cohen MM, et al. Testing internal consistency and construct validity during evaluation of performance in a patient simulator. Anesth Analg 1998; 86:1160–1164.

5. Lawshe CH. A quantitative approach to content validity. Lafayette, IN: Personnel Psychology, Inc., 1976; 563–575.

6. Validity in assessments: content, construct & predictive validity. 27 Apr. 2016. Study.com, Mountain View, CA. Available at: http://study.com/academy/lesson/validity-in-assessments-content-construct-predictivevalidity.html.

7. Cronbach LJ, Meehl PE. Construct validity in psychological tests. S.l.: S.n.; Psychological Bulletin, Washington DC, 1955.