EDUCATION
RESEARCH SERVICE

COMMENTARY

Current state and future directions of the National Board of Chiropractic Examiners

Norman E. Ouzts, Jr, DC, Igor Himelfarb, PhD, Bruce L. Shotts, DC, MS, and Andrew R. Gow, DC, LAc

The objective of this paper is to describe changes made to chiropractic national board examinations in the United States, including methodologies in test scoring, and to discuss future directions in test development and administration being considered by the National Board of Chiropractic Examiners (NBCE). Additionally, this paper serves as an introduction to the articles written by the NBCE staff and published in this issue of the journal. Statistical perspective on the properties of a test are presented, and reasons for the NBCE moving to item response theory for test scoring are described. NBCE consideration of on-demand testing and changes implemented in the Part IV practical examination are also discussed.

Key Indexing Terms: Chiropractic; Educational Measurement; Education; Certification/Standards

J Chiropr Educ 2020;34(1):31–34 DOI 10.7899/JCE-19-24

INTRODUCTION

The objective of this paper is to communicate changes initiated by the National Board of Chiropractic Examiners (NBCE) testing programs and to serve as an introduction to the articles written by NBCE staff that are published in this issue of the journal. In the past 5 years, the NBCE gradually modified statistical models used for scoring, changed the mode of administration for the written examinations, and revised the Part IV exam.

DISCUSSION

The NBCE engages in constant review of its practices and compares its operational procedures of test development, administration, scoring, and reporting to the practices of sister fields in health care. In the last 5 years, the NBCE evaluated all of its products in terms of validity, reliability, fairness, and alignment with the best practices accepted in the health care industry and the field of educational measurement. The first change was that the NBCE implemented computer-based testing (CBT) for the Part I, II, III, and Physiotherapy exams. The CBT administration was previously adopted by the National Board of Medical Examiners,[1] the National Board of Dental Examiners,[2] and the National Board of Osteopathic Medical Examiners.[3]

The second change was to introduce item response theory (IRT) scoring to all NBCE exams. Previously, not all NBCE exams were scored using IRT, which is the practice in the fields of medical,[4–7] dental,[8,9] and osteopathic[10,11] examinations. As a result, a decision was made to gradually adopt new IRT-based scoring methodologies while closely monitoring classification consistency and pass/fail rates.

The third change was to revise the Part IV exam (chiropractic practical exam). The exam needed revisions of the postencounter probe stations in order to better assess the examinees' case management skills. In particular, diagnostic images and laboratory results were excluded from all postencounter probe stations. This was implemented for better evaluation of entry-level competence. Furthermore, electronic presentation of the images was introduced to the diagnostic imaging portion of the exam.

What Is a Test?

As the articles published in this issue address NBCE exams and scoring methodologies, we would like to start by explaining the essential parts of a test. A test (or exam) is an assessment intended to measure examinees' competency in a subject, aptitude, or skill.[12] However, the relationship between a test score and actual knowledge is often misinterpreted or even misunderstood. Because this can be confusing, we provide a definition of a test from a statistical perspective. The relationship between a test score and actual competency is similar to the relationship between an estimate and a true value (parameter). Test scores are sample-based estimates of the real competency. The parameter, if we could measure it directly, is a population-based true value.[13] For example, it may be very

hard to measure exactly how many fish are in a lake, but we can estimate it. One may count fish in a part of the lake and then rescale this estimate to the size of the lake. The precision of such an estimation would depend on the quality of the procedure. Testing is merely a technique that samples the true competency of examinees, while the relationship between the score and reality depends on the quality of a test.

Statistical estimates calculated on a sample are valid only when the sample is representative of the corresponding population. Going back to the fish example, the area of the lake where we collect measurements should be very similar to the rest of the lake. We should not collect our measures next to a bridge where people throw food into the water, thereby attracting the fish, nor should we count the fish in an empty part of the lake.

The same is true for test scores. To make valid inferences, the scores need to be representative of the competency of the test takers. However, some degree of error between the sample-based estimate and the true value is always expected. This error is termed *sampling error*, and our task is to minimize it. The precision of an estimate is inversely related to the degree of sampling error.[14] The topics of score precision, validity, and reliability, as well as the efforts NBCE makes to ensure them, are discussed in Himelfarb[15] and Himelfarb et al.[16]

Classical Test Theory vs Item Response Theory

Classical test theory (CTT) and IRT are 2 major measurement theories employed to model item responses. For many years, the NBCE's scoring procedures were based on CTT, which is known as the *true score theory*.[17] In recent years, innovative scoring methodologies have been developed and adopted by large-scale assessment programs. IRT-based methods are considered contemporary, more informative, and more effective. While the CTT and IRT frameworks are similar in many ways, CTT has received substantial criticism over the past several decades.[18]

In accordance with CTT, every test taker's true score (a score that would be obtained if there were no errors in measurement) is influenced by an additive, unsystematic error term (the difference between the true and observed scores). The implication is that, for individual test takers, the test is an imprecise tool.[18] Furthermore, since the true score is defined using a specific set of items on a test, it is entirely dependent on that set of items. In IRT, on the other hand, individual ability for each test taker is estimated and accounted for when the final score is derived. Another difference between the theories exists in the estimation of the standard error of measurement (the imprecision in the test). With CTT, this error is a constant for all examinees, while in IRT, it could be estimated for different levels of ability. The IRT-based estimation of standard error of measurement is more realistic.[19(p482)]

In 2014, the NBCE began transitioning its exams to IRT scoring. Today, all NBCE prelicensure exams are scored using IRT models. Thus, the numbers provided to test takers, chiropractic institutions, and state licensing boards are calculated based on more realistic assumptions and are more precise. Himelfarb[15] and Himelfarb et al.[16] provide an in-depth discussion of the differences between the 2 theories. For example, Himelfarb et al.[20] explain how Part IV is scored with IRT models using the diagnostic imaging portion of the exam.

Future Directions: Is On-Demand Testing Right for Chiropractic?

Beginning in 2019, the NBCE fully implemented CBT for the Part I, II, III, and Physiotherapy examinations, which allows for more test innovation, more convenient scheduling, and a smaller scoring window. Adopting Parts I and II for CBT administration, we reduced the exams to 300 items each (50 items per domain). Preliminary validity studies have been conducted,[21] and now a validity argument is being built while developing and using the assessments.

Recently, however, there have been several inquiries concerning on-demand testing (a testing service that is available anytime) and its feasibility for the chiropractic profession. Although on-demand testing is being increasingly used in many areas of assessment, it has not been easily adopted by high stakes testing[22] such as the NBCE licensure exam programs. One of the major issues with on-demand testing is that some of the psychometric methods used in conventional testing are no longer available when tests are administrated on demand. While new methodologies have been developed, today they require per-administration sample sizes prohibitively larger than that which the chiropractic profession is currently able to produce.[23] Currently, per-administration sample sizes for written exams are between 1,000 and 1,200 test takers, which includes first-time examinees and repeat test takers. However, the NBCE uses only item responses from a norming group (first-time, nonaccommodated test takers) to fit the IRT models, which further reduces the available sample sizes. Excluding repeaters helps to control for the possible effect of repeaters on equating.[24,25]

If NBCE exams were given on demand to our current number of examinees, we would not have sufficient data to perform psychometric analyses properly, as specified by the best practices detailed in the *Standards for Educational and Psychological Testing*.[26]

Yet, in the context of competency assessment, a computer adaptive testing (CAT) approach may be a conceivable alternative to on-demand tests.[27] CAT is a form of assessment that adapts to the ability level of each examinee. Based on the examinee's previous responses, for subsequent questions CAT selects from test items that maximize the precision of the exam. Consequently, test takers with different ability levels will receive different tests. IRT methodology is used to select optimal items for the test, which are chosen based on the statistical estimates of the information and difficulty. The advantage of using CAT is in uniform precision for all test takers, whereas traditional testing provides the best precision for examinees in the middle of ability range. Matching the difficulty of items on the test with the ability of the test taker allows for obtaining maximum information from each item, so the length of the test could be reduced without loss of reliability. Furthermore, by transitioning to CAT, the NBCE will be able to increase the number of testing windows.

The basic principle behind adaptive testing is simple: avoid asking questions that are much too difficult or much too easy for a particular examinee. It is likely that able examinees will answer easy items correctly and struggling test takers will stumble on hard questions; thus, their responses are not particularly informative. Much more is learned by administering questions that challenge, but don't overwhelm, the examinee. Properly identifying and then presenting these questions is the goal of CAT.

CAT is designed to maximize measurement efficiency, or the precision of test scores in relation to test length. This means that an adaptive test can either save time by being shorter than a conventional test of equal precision or improve score quality by being more precise than a conventional test of equal length.[27] The employment of CAT is a laborious process; however, with the CAT implementation, the majority of examinees will take shorter tests with items more closely selected at their ability levels. For this reason, the NBCE is committed to exploring CAT transition.

Content validity studies using the Delphi method[28] for the Part I and II exams are planned for 2020. During the course of study, the NBCE will provide current test plans and weights to each doctor of chiropractic program, and each will have an opportunity to indicate areas in the test plans that require additions or deletions. The NBCE will summarize the results received from the colleges. The test plans will be modified according to study results.

Future Directions for the Part IV Practical Examination

In January 1996, the NBCE introduced the Part IV exam, the large-scale assessment for chiropractors, which was made up of 3 domains: diagnostic imaging, chiropractic technique, and case management.[29] The main component of Part IV is the objective structured clinical examination (OSCE), an assessment designed to test clinical skill performance and competency by simulating real-world procedures.[30] The OSCE is a standard mode of assessment of medical competency and clinical skills in the United States, Canada, and the United Kingdom.[31–33]

Several changes were made to the Part IV exam. First, diagnostic images and laboratory findings were removed from postencounter probe stations. This was implemented for better evaluation of entry-level competence. In accordance with best practices in imaging and patient care, initial case management of some conditions would not require imaging. Second, electronic presentation of the images (on a computer) was introduced to the diagnostic imaging part of the exam. Finally, all domains of the Part IV exam are now scored using IRT methodology. In 2020, an investigation will begin to determine if there are options to further modernize the Part IV exam and further align the chiropractic OSCE with the standard practices currently accepted in health care.

CONCLUSION

We see testing as a dynamic process that occasionally requires an update. To stay relevant, the NBCE needs to keep up with the modern practices in measurement. Transitioning to IRT was a necessary step in that direction. The implementation of CBT challenged us to construct a fair, valid, and reliable assessment system, to minimize examinees' frustration, and to limit sources of test anxiety. CBT also prompted us to shorten the test and increase the number of testing windows. We hope that the rest of the chiropractic community will share our perception of successful testing on computers. For our part, we will augment the effectiveness of this new mode of assessment through better orientation, easier registration, and possibly, even more testing windows.

Ignoring the evolution in assessment of skills and changes in testing technology may result in a mismatch between the professional skills and testing instruments. The current efforts to modernize the Part IV exam will align the chiropractic OSCE with the standard practices currently accepted in health care. Certainly, a critical part of this transition is to ensure that there will be no disadvantage to our examinees.

FUNDING AND CONFLICTS OF INTEREST

This work was funded internally. The authors have no conflicts of interest to declare relevant to this work.

About the Authors

Norman Ouzts is the chief executive officer of the National Board of Chiropractic Examiners (901 54th Avenue, Greeley, CO 80634; nouzts@nbce.org). Igor Himelfarb is the director of psychometrics and research of the National Board of Chiropractic Examiners (901 54th Avenue, Greeley, CO 80634; ihimelfarb@nbce.org). Bruce Shotts is the director of written exams of the National Board of Chiropractic Examiners (901 54th Avenue, Greeley, CO 80634; bshotts@nbce.org). Andrew R. Gow is the director of practical examinations of the National Board of Chiropractic Examiners (901 54th Avenue, Greeley, CO 80634; agow@nbce.org). Address correspondence to Igor Himelfarb, 901 54th Avenue, Greeley, CO 80634; ihimelfarb@nbce.org. This article was received May 22, 2019, revised May 31, 2019, and accepted November 3, 2019.

REFERENCES

1. National Board of Medical Examiners. Examinee Support Services Web site. https://www.nbme.org/health-profession-services/ESS.html. Updated December 11, 2018. Accessed October 23, 2019.
2. Joint Commission on National Dental Examinations. National Board Dental Examination (NBDE) Part I 2019 Guide. https://www.ada.org/~/media/JCNDE/pdfs/2019_NBDE_PartI_Guide.pdf?la=en. 2019. Cited October 23, 2019.

3. National Board of Osteopathic Medical Examiners. COMPLEX-USA: The pathway to osteopathic medical practice and licensure in the United States Web site. https://www.nbome.org/exams-assessments/comlex-usa/2019. Accessed October 23, 2019.
4. Downing SM. Item response theory: applications of modern test theory in medical education. *Med Educ*. 2003;37:739–745.
5. Hambleton RK, Slater SC. Item response theory models and testing practices: current international status and future directions. *Eur J Psychol Assess*. 1997;13(1):21–28.
6. Luecht RM, Nungester RJ. Some practical examples of computer-adaptive sequential testing. *J Educ Meas*. 2005;35(3):229–249.
7. National Board of Medical Examiners. Scoring Services Web site. https://www.nbme.org/health-profession-services/scoring.html. 2019. Accessed October 23, 2019.
8. Neumann LM, MacNeil RL. Revisiting the National Board Dental Examination. *J Dent Educ*. 2007;71(10): 1281–1292.
9. Yang C-L, Neumann LM, Kramer GA. Assessing context effects on test validity of the National Board Dental Examination Part I. *J Dent Educ*. 2011;76(4): 395–406.
10. Shen L. Constructing a measure for longitudinal medical achievement studies by the Rasch model one-step equating. Paper presented at: American Educational Research Association; April 1993; Atlanta, GA.
11. Larger MM, Swanson DB. Practical considerations in equating progress tests. *Med Teach*. 2010;32:509–512.
12. Anastasi A. *Psychological Testing*. 6th ed. New York: Macmillan; 1988.
13. Freedman D, Pisani R, Purves R. *Statistics*. New York: WW Norton and Company; 2007.
14. Crano WD, Brewer MB, Lac A. *Principles and Methods of Social Research*. 3rd ed. New York: Routledge; 2014.
15. Himelfarb I. A primer on standardized testing: History, measurement, classical test theory, item response theory, and equating. *J Chiropr Educ*. 2019; 33(2):151–163.
16. Himelfarb I, Shotts B, Tang N-E, Smith M. Score production and quantitative methods used by the National Board of Chiropractic Examiners for post-exam analyses. *J Chiropr Educ*. 2020;34(1):35–42.
17. Crocker L, Algina J. *Introduction to Classical and Modern Test Theory*. New York, NY: Harcourt, Brace Jovanovich College Publishers; 1986.
18. Raykov T, Marcoulides G. On the relationship between classical test theory and item response theory: from one to the other and back. *Educ Psychol Meas*. 2016;76(2):325–338.
19. Mango C. Demonstrating the differences between classical test theory and item response theory using derived test data. *Int J Educ Psychol Assess*. 2009;1(1): 1–11.
20. Himelfarb I, Seron MA, Hyland JK, et al. The transition to digital presentation of the diagnostic imaging domain of the part IV examination of the National Board of Chiropractic Examiners. *J Chiropr Educ*. 2020;34(1):52–67.
21. Himelfarb I. Examining the accuracy of classification of a redesigned exam: decision consistency studies. Proceedings of the 10th World Federation of Chiropractic/Association of Chiropractic Colleges Education Conference; October 24, 2018; London, England. *J Chiropr Educ*. 2018;32(2):e181
22. He Q. On-demand testing and maintaining standards for general qualifications in the UK using item response theory: possibilities and challenges. *Educ Res*. 2012;54(1):89–112.
23. Hambleton RK, Swaminathan H. *Item Response Theory: Principles and Applications*. Boston: Kluwer; 1996.
24. Kim S, Walker ME. Effect of repeaters on score equating in a large-scale licensure test. Report number ETS RR-09-27. Princeton, NJ: Educational Testing Service; 2009.
25. Yang W-L, Bontya AM, Moses TP. Repeater effects on score equating for a graduate admissions exam. Report number ETS RR-11-17. Princeton, NJ: Educational Testing Service; 2011.
26. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: The Associations and Council; 2014.
27. Davey T. *A Guide to Computer Adaptive Testing Systems*. Washington, DC: Council of Chief State School Officers; 2011.
28. Dalkey N, Helmer O. An experimental application of the Delphi method to the use of experts. *Manag Sci*. 1963;9(3):458–467.
29. Townsend PD, Christensen M, Kreiter CD, zumBrunnen JR. Investigating the use of written and performance-based testing to summarize competence on the case management component of the NBCE Part IV–national practical examination. *Teach Learn Med*. 2010;22(1):16–21.
30. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. *Br Med J*. 1975;(1):447–451.
31. Hodges B. OSCE! Variations on a theme by Harden. *Med Educ*. 2003. 37(12):1134–1140.
32. Jain SS, DeLisa JA, Eyles MY, Nadler S, Kirshblum S, Smith A. Further experience in development of an objective structured clinical examination for physical medicine and rehabilitation residents. *Am J Phys Med Rehabil*. 1998.77(4):306–310
33. Zayyan M. Objective structured clinical examination: the assessment of choice. *Oman Med J*. 2011;26(4):219–222.