EDUCATION
RESEARCH SERVICE

ORIGINAL ARTICLE

Score production and quantitative methods used by the National Board of Chiropractic Examiners for postexam analyses

Igor Himelfarb, PhD, Bruce L. Shotts, DC, Nai-En Tang, PhD, and Margaret Smith

Objective: The National Board of Chiropractic Examiners (NBCE) uses a robust system for data analysis. The aim of this work is to introduce the reader to the process of score production and the quantitative methods used by the psychometrician and data analysts of the NBCE.
Methods: The NBCE employs data validation, diagnostic analyses, and item response theory–based modeling of responses to estimate test takers' abilities and item-related parameters. For this article, the authors generated 1303 synthetic item responses to 20 multiple-choice items with 4 response options to each item. These data were used to illustrate and explain the processes of data validation, diagnostic item analysis, and item calibration based on item response theory.
Results: The diagnostic item analysis is presented for items 1 and 5 of the data set. The 3-parameter logistic item response theory model was used for calibration. Numerical and graphical results are presented and discussed.
Conclusion: Demands for data-driven decision making and evidence-based effectiveness create a need for objective measures to be used in educational program reviews and evaluations. Standardized test scores are often included in that array of objective measures. With this article, we offer transparency of score production used for NBCE testing.

Key Indexing Terms: Chiropractic; Education; Psychometrics; Scoring Methods; Statistical Data Analysis

J Chiropr Educ 2020;34(1):35–42 DOI 10.7899/JCE-18-27

INTRODUCTION

According to the Standards for Educational and Psychological Testing, assessment is among the most important contributions of cognitive and behavioral sciences to our society.[1] Decision processes in health care professional testing are often complex and ongoing and pose additional challenges due to numerous regulations and their enforcement by the agencies responsible for safety of the general public.[2]

The National Board of Chiropractic Examiners (NBCE) adheres to the Standards for Educational and Psychological Testing.[1] These principles dictate that the NBCE provide accurate, fair, valid, and reliable assessment results to the intended score recipients, follow specific guidance in assessment development, obey psychometric standards, and respect the rights and responsibilities of the test takers. The goal of the NBCE is to produce scores that are valid and reliable for all test takers and are comparable over time and across test forms. For that purpose, the NBCE follows an established protocol that ensures that these goals are met.

Therefore, our objective for this article is to introduce chiropractic educators to the field of measurement and to demystify the complex and laborious process of scoring NBCE exams. This article provides an overview of principle concepts and modern-day best practices accepted in testing. This is followed by a description and illustration of the multiphase process of score development used by the NBCE using a generated data set that mimics the data structure of NBCE Part I and Part II exams.

OVERVIEW OF TESTING CONCEPTS

Data Validity

Tabachnick and Fidell[3] supplied a checklist for screening data prior to statistical analysis: (1) inspect univariate descriptive statistics for accuracy of input, (2) evaluate the amount and distribution of missing data and deal with the problem, (3) check pairwise plots for nonlinearity and heteroscedasticity, (4) identify and deal with nonnormal variables and univariate outliers, (5) identify and deal with multivariate outliers, and (6) evaluate variables for multicollinearity and singularity. The NBCE follows their suggestions in our data screening procedures.

Furthermore, the NBCE understands the importance of ensuring the accuracy of data used in the process of test score production. After receiving the data from test sites, psychometric data analysts closely examine the match of the data set to the data map, ensuring that the responses for both paper-based and computer-based forms are within the expected ranges.

As part of the data validation procedure, the NBCE psychometric team, in collaboration with the Written Examinations and Part IV (Practical Testing) staff, established criteria to determine whether examinees exhibit a valid attempt to respond to the items on tests. For paper-based testing, we expect the test takers to mark at least some questions in the answer document. For the computer-based testing, we inspect the examinee response data, along with timing information, and determine whether a test taker made a valid attempt on the test. In addition, a score for a test or part of a test may be invalidated if an examinee completes a section in a very short time and/or receives a very low score. Although we flag item responses that do not meet valid attempt criteria, score invalidations are extremely rare.

Recent research[4,5] suggested the use of item response theory (IRT) to identify the cases that behave inconsistently with model assumptions. IRT model-based fit indices may serve as validity estimates for cases with particular response patterns.[6] The NBCE uses IRT-based modeling to establish data validity and conduct forensic data analyses.

Missing Data

Statistical analyses involving inference and prediction become problematic in the presence of missing data.[7,8] Several methods of dealing with missing responses in psychological and educational research have been identified and developed.[9] However, further consideration should be given to this issue in operational psychometrics, as the effects of missing data on the estimation accuracy of IRT parameters is well documented.[10]

Test takers may omit responses when a page layout is complicated or when they have to follow a long passage or task and do not respond due to fatigue or intimidation. Missing data can also occur when test takers run out of time or are unmotivated, overly anxious, fatigued, or overwhelmed.[11] The NBCE is closely monitoring cases with missing data as well as investigating the possible causes of missing data from each testing administration.

Scoring

Scoring could be defined as converting raw item responses to scored responses according to a rubric that makes this process a function of the item type. Items on a test are usually classified as selected-response (SR) or constructed-response (CR) items. In this study, we only review the scoring process for SR items, where the examinees select a correct answer from a limited number of choices. Examples of SR items include multiple-choice items, true-or-false items, or matching items. Scoring for SR items is called objective scoring, which means that no judgment is required for raters to score an item; thus, scored items will have the same score regardless of who scores them. The majority of SR items are scored using the "number correct method," where the overall test score is the total number of correct responses.[12] For example, consider the following test item:

What is the capital of France?

A. London

B. San Francisco

C. Madrid

D. Paris

The item is a multiple-choice item that contains a single correct answer: Paris. This item is scored dichotomously in the following way: 1 = if the response is Paris, 0 = if otherwise. Other scoring methods are available for test items with more complex designs.[13,14]

Diagnostic Item Analysis

The operational psychometric procedures include an evaluation of examinees' performance as well as the performance of items on the test. The first step in this evaluation is to conduct a diagnostic item analysis (DIA), which is a statistical analysis based on classical test theory (CTT).[15] The DIA provides measurement and bias information about items. This information is used for item reviews, test construction and revisions, technical reports, and other psychometric documentation. The DIA shows the number and percent of test takers responding to each answer choice, the p values, point-biserial correlations, and other useful statistics. Further descriptions of these statistics follow.

Item Difficulty

Item difficulty is defined as the proportion of examinees who answered the item correctly, also known as the p value. The formula for p value is

$$p = \frac{N_{ic}}{N_i}$$

where N_{ic} is the number of examinees who answered item i correctly and N_i is the total number of examinees who attempted the item.

Commonly, for dichotomously scored items, the difficulty of an item is measured by the proportion of test takers who answered the question correctly. The range of proportion correct is 0 to 1, with 0 indicating that all examinees responded to an item incorrectly and 1 indicating that all examinees responded to an item correctly. Higher p values indicate easier items and/or more able populations. Desired p values generally fall within the range of .25 to .95. For multiple choice items, Thompson and Levitow[16] suggested that the ideal difficulty for an item is slightly higher than the middle point between the percentage of answering correctly by guessing (25% for the 4-option multiple-choice items) and all examinees answering correctly (100%). For polytomously scored items, the p value represents the average item score or the proportion of the maximum obtainable score.[17]

Desired values generally fall within the range of 30% to 80% of the maximum obtainable score.

Item Discrimination

Item discrimination refers to the extent that a test item distinguishes between examinees with different levels of ability. The index of item discrimination is derived using correlation. The foundation of the correlation-based approach is the Pearson product-moment correlation used to measure the strength of linear relationship between 2 normally distributed variables.[18] The formula for the Pearson product-moment correlation is

$$r_{x,y} = \frac{COV(x,y)}{SD_x SD_y}$$

where $COV(x, y)$ is the covariance between variables x and y, SD_x is the standard deviation for x, and SD_y is the standard deviation for y.

When measuring item discrimination for dichotomously scored items, a special case of the Pearson product-moment correlation, called point-biserial correlation, is used.[11] The formula for the point-biserial correlation is

$$r_{pbis} = \frac{\bar{X}_s - \bar{X}_\mu}{S_Y}\sqrt{\frac{p}{q}}$$

where $\bar{X}_s$ is the mean test score for examinees who provide a correct response to the item, $\bar{X}_\mu$ is the mean test score for examinees who responded to the item incorrectly, S_Y is the standard deviation of the test score, p is the proportion of examinees who respond to the item correctly, and q is the proportion of examinees who responded to the item incorrectly.

An item is considered to perform well if high-ability test takers tend to answer correctly and low-ability test takers tend to answer incorrectly. An item with negative or extremely low correlations indicates serious problems and should be reviewed.

Distractor Analysis

Analysis of distractors is required in determining the usefulness of the attractiveness of each option. A distractor should be a plausible choice, reflecting a common misconception. If a distractor fails to attract examinees with lower ability levels, the response option should be modified. A discrimination index (eg, point-biserial correlation coefficient) should also be calculated for each distractor to determine whether it is performing correctly.[19] We expect the discrimination index for distractors to be zero or negative.

IRT

The logic behind testing is to develop an instrument that, with a number of items, will reliably measure the ability of interest. Then, using the pattern of item responses, with reasonable precision, the place on the ability continuum for each examinee can be determined. In the IRT literature, the ability parameter is denoted θ_j. Then, using mathematical models, a probability of responding correctly on a test item conditional on ability could be calculated. For a properly functioning test item, this probability will be near 0 for examinees who are low on the ability continuum and near 1 for examinees who are high on the continuum. The S-shape curve connecting the ability (x-axis) and the probability of the correct response (y-axis) for a particular item is known as the item characteristic curve.[20]

CTT is built around the framework of linear models—it uses the linear decomposition to separate the true score from error. IRT models relate the student's ability to item scores using a nonlinear framework.[21] The IRT models for dichotomous data differ in the number of parameters included in the models. The 1-parameter logistic model (1PL) estimates the probability of the correct response to an item as a function of the difference between the examinee's ability and item difficulty. This estimation becomes possible because the examinee's ability and item difficulty are on the same, logit scale. The logit, or the log odds, is a logarithm of odds $\frac{p}{1-p}$, where p is the probability for the event of interest. Thus, $logit(p) = \log\left(\frac{p}{1-p}\right)$. A unit increase on the logit scale represents an increase in odds of the correct response to the item. Logit scale is a perfect representation of the interval scale,[22] which is considered to be one of the key advantages when using logit. Furthermore, the logarithmation allows creating a continuum between the dichotomously scored responses.

The 2-parameter logistic model (2PL) introduces a discrimination parameter that is the slope of the curve—the steeper the slope, the better the discrimination. The model is advantageous when compared to 1PL due to a closer fit to the response data resulting in more precise parameter estimates. Finally, the 3-parameter logistic model (3PL) introduces the guessing parameter in addition to the difficulty and discrimination estimated by the 2PL, which is the probability of providing a correct response to an item by chance. The value of the guessing parameter does not vary as the function of ability level.[20]

Calibration

Calibration is a process of fitting IRT models to scored item responses. The purpose of item calibration is to obtain IRT parameters (difficulty, discrimination, and guessing) for each item on a test. Himelfarb[23] detailed the history, assumptions, and models of IRT. The NBCE uses the 3PL IRT model[24,25] for dichotomously scored items and the graded response model[26] or the generalized partial credit model[27] for items scored using more than 1 correct answer. The calibration of item responses provides us with many features that are helpful in the decision-making process about test takers and test items. In a later section, we will discuss the item-related information that the NBCE obtains from calibration procedures. The following illustrates the operational psychometric procedures that the NBCE employs.

SIMULATED DATA SET AND SCORE DEVELOPMENT

The purpose of this study is to introduce the reader to the course of test score production. To illustrate the

Table 1 - Raw Item Responses

	Item										
Examinee	1	2	2	4	5	6	7	8	9		20
1	1	2	1	4	3	2	1	3	4	:	2
2	1	3	2	4	1	4	1	3	1	:	2
3	1	3	3	2	1	1	1	3	4	:	2
4	1	3	3	4	4	1	1	3	4	:	3
5	1	3	3	4	1	3	1	3	4	:	1
6	4	4	1	3	3	3	1	3	4	:	1
:	:	:	:	:	:	:	:	:	:	:	:
1303	3	2	2	2	1	2	3	3	2	:	2

scoring, 1303 synthetic item responses to 20 multiple-choice (MC) items with 4 response options to each item were generated. The "psych" package[28] within R programing language[29] version 3.3.1 was used to generate unidimensional item responses. It was assumed that each item has only 1 correct answer. Throughout the remainder of the article, we will refer to these generated data as "the Test."

A popular scoring schema for MC items is dichotomous scoring. When scored dichotomously, an examinee receives 1 scored point for selecting the correct response and 0 scored points for choosing a distractor.[30] Tables 1 and 2 present raw and scored item responses for the Test, respectively.

The DIA based on CTT methods was performed using raw data responses on the Test. We used the Structured Query Language[31] to generate the DIA. The estimated statistics included a number of examinees choosing a particular response category, a *p* value, and an estimate of point-biserial correlation for each response category on an item.

The IRT calibration was performed with the 3PL IRT model.[32] The 3PL is a model that estimates the parameters of item discrimination (a_i) and item difficulty (β_i) with an additional parameter, γ_i—the lower asymptote of the item characteristic curve, representing the probability of a test taker with a low ability providing a correct answer to an item *i*. The inclusion of this parameter suggests that test takers who score low on the latent trait may still provide a correct response by chance. This parameter is referred to as "guessing." The following is the mathematical representation of the 3PL IRT model:

Table 2 - Scored Item Responses

	Item										
Examinee	1	2	2	4	5	6	7	8	9		20
1	1	0	0	1	1	1	1	0	1	:	0
2	1	1	0	1	1	1	0	0	1	:	0
3	1	0	1	1	1	0	1	0	1	:	0
4	1	1	1	1	0	0	0	0	0	:	1
5	1	0	1	1	0	0	1	0	1	:	0
6	0	1	1	1	1	1	1	1	1	:	1
:	:	:	:	:	:	:	:	:	:	:	:
1303	1	1	1	1	1	1	0	0	1	:	1

Table 3 - Diagnostic Item Analysis for Item 1

Response Category	n	*p*	*rp-b*	Lower	Mid 50%	Mid 75%	Upper
A*	851	0.78	0.27	0.19	0.33	0.25	0.24
B	38	0.03	−0.16	0.37	0.50	0.11	0.03
C	138	0.13	−0.37	0.54	0.32	0.13	0.01
D	64	0.06	−0.30	0.59	0.27	0.14	0.00

$$P(u_i = 1|\ \theta, \alpha, \beta, \gamma) = \gamma_i + (1 - \gamma_i)\frac{e^{Da_i(\theta_j - \beta_i)}}{1 + e^{Da_i(\theta_j - \beta_i)}}$$

Table 3 presents the DIA results for item 1 on the Test. The key, A, is evidently preferred by the majority of test takers (n = 851). The *p* value = .78, which demonstrates that this item is in the acceptable difficulty range. The item-total correlation for the key is positive, $r = .29$, while the correlations for distractors are all negative. Figure 1 presents smoothed plots of the key and each of the distractors on item 1. The x-axis represents 4 categories of scores: lower, middle 50%, middle 75%, and upper. The lower category is the proportion of test takers from the lowest score group choosing the response, while upper is the proportion from the highest score group. The y-axis represents the percentage of test takers in each category who chose that response. The graph shows that test takers of lower category are not able to differentiate between the key and the distractors. For the middle 50%, the distractor B is preferred over the key. However, for the middle 75% and upper score categories, there is a clear prevalence of the key over the distractors. Based on the numerical and graphical analyses, we conclude that item 1 is performing appropriately.

Table 4 presents the DIA for item 5 on the Test. Similar to item 1, the key is preferred by the majority of the test takers (n = 844). The item-total correlation for the key is positive and within the accepted range, while the correlations for the distractors are all negative. The numerical analysis advises that A, the distractor, is attracting almost 3 times as many test takers as option B or D. Figure 2 shows that the key is visibly prevalent for examinees in the middle 75% and upper score categories.

For calibration, Table 5 displays the item-parameter estimates for *a* (discrimination), *b* (difficulty), and *c* (guessing) and the standard errors associated with these estimations obtained via calibration of the Test using the 3PL IRT model. The item difficulty represents the point on the ability scale where a test taker has a 50% probability (point of median probability) of providing a correct response to the item. The accepted range for difficulty is

Table 4 - Diagnostic Item Analysis for Item 5

Response Category	n	*p*	*rp-b*	Lower	Mid 50%	Mid 75%	Upper
A	145	0.13	−0.41	0.57	0.34	0.07	0.02
B	52	0.05	−0.30	0.67	0.29	0.04	0.00
C*	844	0.77	0.36	0.16	0.33	0.27	0.24
D	50	0.05	−0.27	0.62	0.32	0.06	0.00

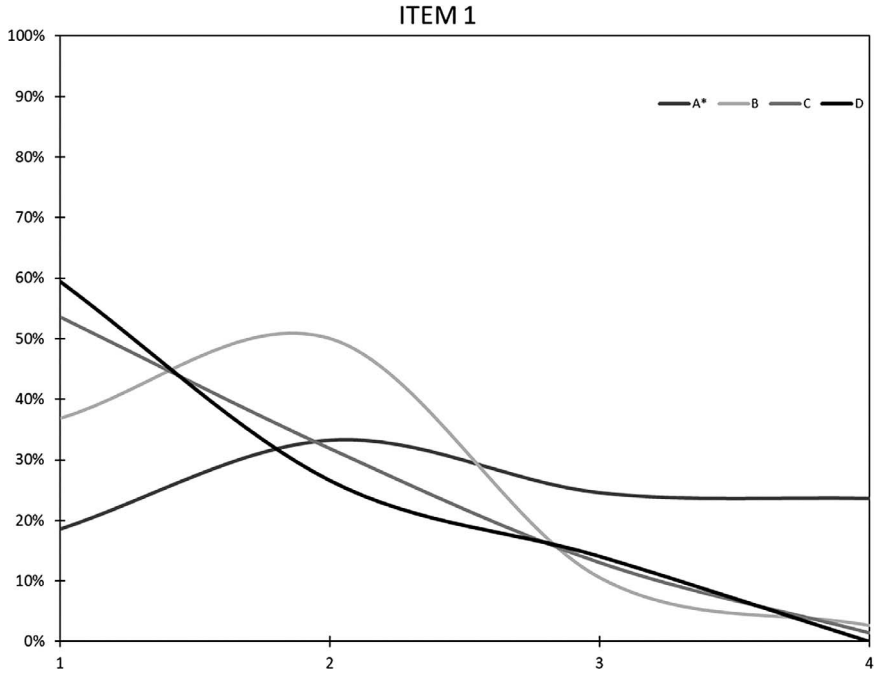

Figure 1 - Diagnostic item analysis, item 1 on the Test; 1 = lower, 2 = mid 50%, 3 = mid 75%, 4 = upper.

between −4.0 and 4.0; however, items with values above +2.0 are considered hard, and items with values below −2.0 are considered easy. The discrimination parameters are not limited in range; however, negatively discriminating items are discarded from a test.[32] The estimates of guessing signify pseudochance; these are the values of the asymptote for the curve representing an item.

Figure 3a demonstrates the item characteristic curves for the 20 items on the test and Figure 3b the item information functions. The x-axis on both graphs represents the test takers' ability or the latent trait, while the y-axis on the left graph shows the probability of the correct response conditional on ability level. On the right graph, the y-axis represents the amount of information each item provides for a specific ability level.

The items to the right on both graphs are more difficult, while items on the left are easier. On the left graph, the items with the steeper slopes are better-discriminating

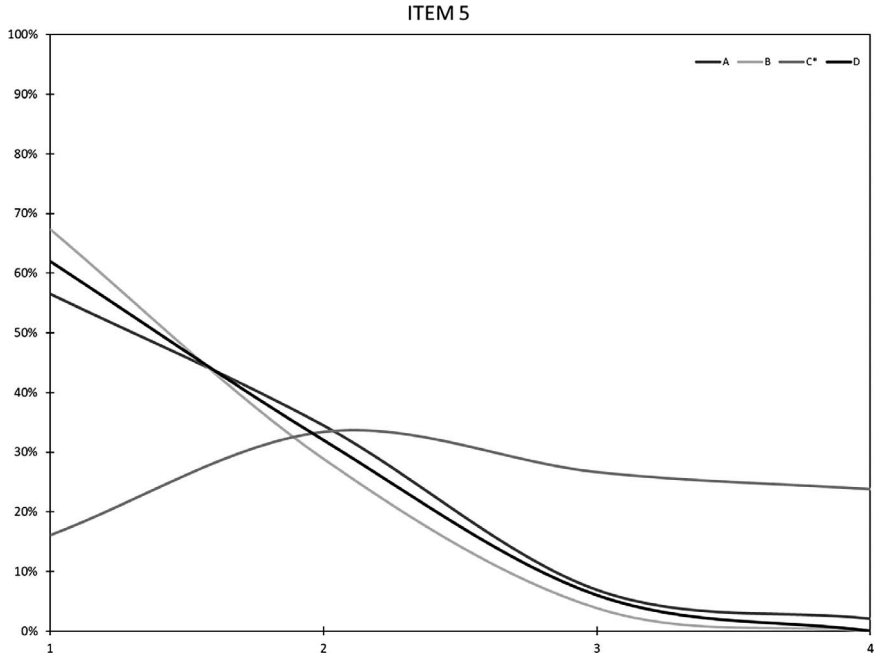

Figure 2 - Diagnostic item analysis, item 5 on the Test; 1 = lower, 2 = mid 50%, 3 = mid 75%, 4 = upper.

Table 5 - Item-Parameter Estimates, 3-Parameter Logistic Model

Item	*a*	SE *a*	*b*	SE *b*	*c*	SE *c*
1	1.02	0.19	−3.48	0.52	0.01	0.11
2	0.73	0.15	−1.76	0.98	0.03	0.36
3	0.90	0.12	−2.13	0.24	0.00	0.03
4	0.93	0.53	−0.62	1.54	0.69	0.20
5	1.62	0.72	1.30	0.16	0.31	0.05
6	0.69	0.24	−1.50	1.73	0.24	0.46
7	1.21	0.15	−1.78	0.21	0.01	0.09
8	2.00	0.64	0.20	0.19	0.40	0.07
9	1.40	0.19	−2.45	0.23	0.00	0.04
10	1.15	0.45	−1.79	1.62	0.17	0.79
11	1.08	0.28	−0.79	0.60	0.09	0.25
12	1.98	0.99	−0.51	0.56	0.82	0.07
13	1.83	0.91	1.64	0.22	0.31	0.03
14	0.89	0.38	0.24	0.66	0.16	0.21
15	1.29	0.38	−1.71	1.01	0.18	0.53
16	2.23	0.85	−1.09	0.51	0.78	0.10
17	0.73	0.11	−1.98	0.41	0.01	0.13
18	1.04	0.37	−0.45	0.74	0.32	0.22
19	0.94	0.27	−0.66	0.73	0.21	0.24
20	1.02	0.12	−1.60	0.20	0.01	0.07

items. From the numerical information and the graphs, item 1 appears to be the easiest on the Test, while item 13 is the hardest. On the right graph, the items with higher curves provide more information regarding ability. This plot helps to see which item is more informative for each ability segment.

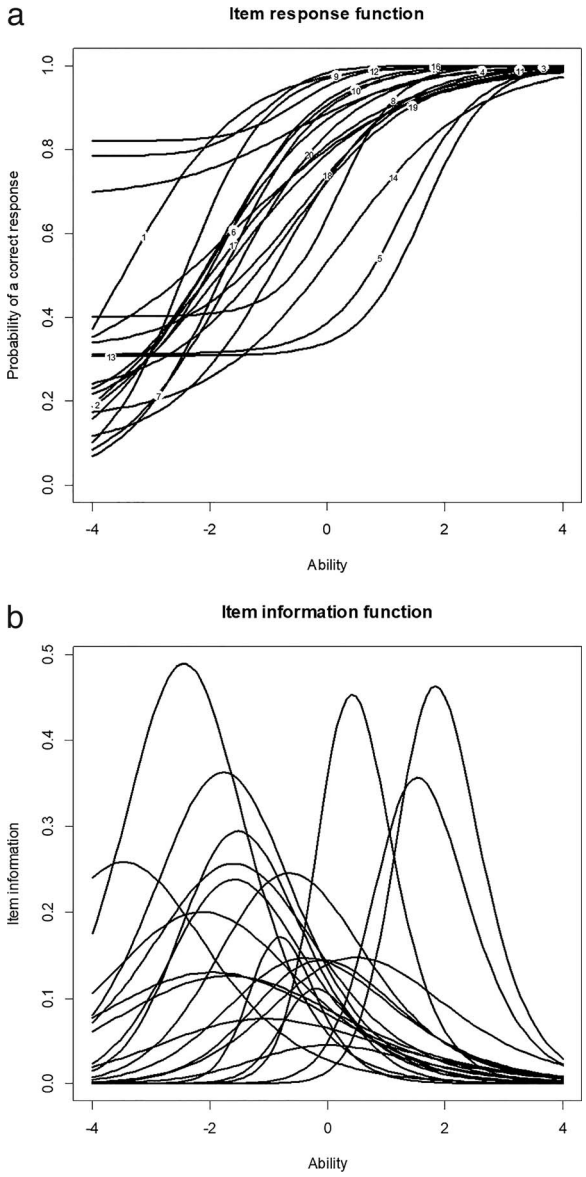

Figure 3a - Item characteristic curves for the Test.
Figure 3b - Information functions for the Test.

DISCUSSION

Professional assessment is a key component for any evidence-based practice.[33] According to the Standards Educational and Psychological Testing,[1] assessment is intended to provide the public, including employers, and governing agencies with a dependable mechanism for identifying practitioners who have met particular requirements and are ready to practice according to established standards. To be able to provide stakeholders with that mechanism, professional testing programs must ensure a close connection between the occupation and the content of the test. The process of gathering such evidence is called validation and is never ending— the use of test scores may be valid for one purpose and not valid for another. Validity would not be possible without reliability,[34] which is a quantification of measurement precision for test scores.[35] The theory of reliability assumes that every examinee possesses a latent true score—the true parameter indicating the degree of knowledge he or she truly has. A test is then an inference that provides an estimate of that parameter.

Error is an inherent factor of measurement. The process of classification is always susceptible to type I error (the probability of rejecting the null hypothesis when null is true) and type II error (the probability of not rejecting the null hypothesis when null is false). The sources for type I and type II errors are countless and may include test taker–related factors such as fatigue or anxiety and psychological or situational factors. However, the goal of a testing program is to minimize the errors related to the instrument (test) by striving to increase the validity and reliability.

To be able to produce a test score, a scoring model needs to be assumed. Every model is a simplification of reality. For example, when a child learns that 1 apple plus another apple equals 2 apples, this models the higher mathematical concept of addition. Thus, often, the simplification of reality is quite useful, as it provides explanation of a phenomenon and affords the ability for prediction. As Gorge Edward Pelham Box, a great British statistician, once said, "All models are wrong, but some are useful."[36]

CONCLUSION

Today the faculty and educational administrators in all sectors of American higher education follow high-stakes accountability policies. The demands for data-driven decision making and evidence-based effectiveness create a need for objective measures to be used in educational program reviews and evaluations. Standardized test scores are often included in that array of objective measures. Thus, educational institutions may base faculty members' evaluations on how well their students do on standardized tests, which leads to implications for professional development, compensation, benefits, and tenure.[37] In turn, faculty members express their frustration criticizing the validity and reliability of scores and the legitimacy of agencies that produce these scores. With this article, we offer transparency of score production and hope that faculty members will take time to understand the seriousness of the work involved.

ACKNOWLEDGMENT

The authors would like to thank Alison Day for her review and editing contributions of this article.

FUNDING AND CONFLICT OF INTEREST

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

About the Authors

Igor Himelfarb is the director of the Department of Psychometrics and Research at the National Board of Chiropractic Examiners (901 54th Avenue, Greeley, CO 80634; ihimelfarb@nbce.org). Bruce L. Shotts is the director of written examinations at the National Board of Chiropractic Examiners (901 54th Avenue, Greeley, CO 80634; bshotts@nbce.org). Nai-En Tang is a psychometric data analyst in the Department of Psychometrics and Research at the National Board of Chiropractic Examiners (901 54th Avenue, Greeley, CO 80634; ntang@nbce.org). Margaret Smith is a senior data analyst in the Department of Psychometrics and Research at the National Board of Chiropractic Examiners (901 54th Avenue, Greeley, CO 80634; msmith@nbce.org). Address correspondence to Igor Himelfarb, 901 54th Avenue, Greeley, CO 80634; ihimelfarb@nbce.org.

This article was received September 16, 2018; revised December 20, 2018; and accepted January 18, 2019.

Author Contributions

Concept development: IH. Design: IH. Supervision: BS. Data collection/processing: IH, NET, MS. Analysis/interpretation: IH, MS. Literature search: NET. Writing: IH, BS. Critical review: IH, BS.

REFERENCES

1. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education; 2014.
2. Grim K, Rosenberg D, Svedberg P, Schon UK. Development and usability testing of a web-based decision support for users and health professionals in psychiatric services. *Psychiatr Rehabil J*. 2017;40(3):293–302.
3. Tabachnick BG, Fidell LS. *Using Multivariate Statistics*. 7th ed. Boston, MA: Pearson; 2018.
4. Falk CF, Cai L. A flexible full-information approach to the modeling of response styles. *Psychol Methods*. 2016;21:328–347.
5. Huang HY. Mixture random-effect IRT models for controlling extreme response style on rating scales. *Front Psychol*. 2016;7:1–16.
6. DiStefano C, Liu J, Greer F. Identification of questionable data using validity indices and item response theory methods: examinations with a teacher-rating scale. *Psychol Assess*. 2018;30(4):500–511.
7. Rubin DB, Little RJA. *Statistical Analysis with Missing Data*. 2nd ed. New York, NY: Wiley; 2002.
8. Graham JW. Missing data analysis: making it work in the real world. *Annu Rev Psychol*. 2009;60:549–576.
9. Pigott TD. A review of methods for missing data. *Educ Res Eval*. 2001;7(4):353–383.
10. Wang S, Harris G. Effect of missing data in computerized adaptive testing on accuracy of item parameter estimation: a comparison of NWEA and WINSTEPS item parameter calibration procedures. Paper presented at: American Educational Research Association (AERA); April 26, 2012; Vancouver, BC, Canada.
11. Kerlinger FN, Lee HB. *Foundations of Behavioral Research*. 4th ed. Boston, MA: Thomson Learning; 2000.
12. Hogan TP. *Psychological Testing: A Practical Introduction*. Hoboken, NJ: John Wiley & Sons; 2003.
13. Fray RB. Formula scoring of multiple-choice tests (correction for guessing). *Educ Meas Issues Pract*. 1988;7(2):33–38.
14. Wolf MK, Guzman-Orth D, Lopez A, Castellano K, Himelfarb I, Tsutagawa F. Integrating scaffolding strategies into technology enhanced assessments of English learners: task types and measurement models. *Educ Assess*. 2016;21(3):157–175.
15. Crocker L, Algina J. *Introduction to Classical and Modern Test Theory*. San Diego, CA: Harcourt Brace Jovanovich College Publishers; 1986.
16. Thompson B, Levitow JE. Using microcomputers to score and evaluate test items. *Collegiate Microcomputer*. 1985;3:163–168.
17. Kolen MJ, Brennan RL. *Test Equating, Scaling, and Linking: Methods and Practices*. New York, NY: Springer; 2014.

18. Cohen J, Cohen P, West SG, Aiken LR. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. 3rd ed. New York, NY: Psychology Press; 2002.
19. Millman J, Green J. The specification and development of tests of achievement and ability. In: *Educational Measurement*. 3rd ed. Phoenix, AZ: Oryx Press; 1993: 335–366.
20. Baker FB. *The Basics of Item Response Theory*. 2nd ed. Washington, DC: ERIC Clearinghouse on Assessment and Evaluation; 2001.
21. Yen WM, Fitzpatrick AR. Item response theory. In: *Educational Measurement*. 4th ed. Westport, CT: American Council on Education; 2006:111–153.
22. Stevens SS. On the theory of scales measurement. *Science*. 1946;103(2684):677–680.
23. 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.
24. Brinbaum DL. Some latent trait models and their use in inferring an examinee's ability. In: *Statistical Theories of Mental Test Scores*. Reading, MA: Addison Wesley; 1968:397–460.
25. Lord FM. A*pplications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1980.
26. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monogr Suppl*. 1969;17:1–100.
27. Muraki E. A generalized partial credit model: application of an E-M algorithm. *Appl Psychol Meas*. 1992;16: 159–176.
28. Revelle W. *psych: Procedures for Psychological, Psychometric, and Personality Research* [software]. Version 1.8.12. Evanston, IL: Northwestern University; 2018.
29. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.r-project.org.
30. Diedenhofen B, Musch J. A web interface and R package for the statistical comparison of Cronbach's alpha coefficients. *Int J Internet Sci*. 2016;11:51–60.
31. Chamberlin DD, Boyce RF. Structured Query Language. http://www.iso.org. 2016.
32. Hambleton RK, Swaminathan H. *Item Response Theory: Principles and Applications*. Boston, MA: Kluwer; 1996.
33. Towne TL, De Young KP, Anderson DA. Trends in professionals' use of eating disorder assessment instruments. *Prof Psychol Res Pract*. 2017;48(4):243–250.
34. Brennan RL. Perspectives on the evolution and future of educational measurement. In: *Educational Measurement*. 4th ed. Washington, DC: American Council on Education; 2006:1–17.
35. Haertel EH. Reliability. In: *Educational Measurement*. 4th ed. Washington, DC: American Council on Education; 2006:65–111.
36. Box GEP. Science and statistics. *J Am Stat Assoc*. 1976;71:791–9.
37. Braun HI. *Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models*. Princeton, NJ: Educational Testing Service; 2005. Policy Information Perspective. Report No. 730194.