EDUCATION
RESEARCH SERVICE

ORIGINAL ARTICLE

The transition to digital presentation of the diagnostic imaging domain of the Part IV examination of the National Board of Chiropractic Examiners

Igor Himelfarb, PhD, Margaret A. Seron, DC, John K. Hyland, DC, MPH, Andrew R. Gow, DC, Nai-En Tang, PhD, Meghan Dukes, DC, MSPT, Margaret Smith, and Michele Fisher

Objective: This article introduces changes made to the diagnostic imaging (DIM) domain of the Part IV of the National Board of Chiropractic Examiners examination and evaluates the effects of these changes in terms of item functioning and examinee performance.
Methods: To evaluate item function, classical test theory and item response theory (IRT) methods were employed. Classical statistics were used for the assessment of item difficulty and the relation to the total test score. Item difficulties along with item discrimination were calculated using IRT. We also studied the decision accuracy of the redesigned DIM domain.
Results: The diagnostic item analysis revealed similarity in item function across test forms and across administrations. The IRT models found a reasonable fit to the data. The averages of the IRT parameters were similar across test forms and across administrations. The classification of test takers into ability (theta) categories was consistent across groups (both norming and all examinees), across all test forms, and across administrations.
Conclusion: This research signifies a first step in the evaluation of the transition to digital DIM high-stakes assessments. We hope that this study will spur further research into evaluations of the ability to interpret radiographic images. In addition, we hope that the results prove to be useful for chiropractic faculty, chiropractic students, and the users of Part IV scores.

Key Indexing Terms: Chiropractic; Educational Measurement; Diagnostic Imaging; Psychometrics

J Chiropr Educ 2020;34(1):52–67 DOI 10.7899/JCE-19-2

INTRODUCTION

All jurisdictions in the United States require proficiency in radiography within a chiropractor's scope of practice. Most also permit licensed chiropractors to order and evaluate the reported results of advanced imaging procedures, such as spinal computed tomography scans and magnetic resonance images. The diagnostic imaging (DIM) component of the National Board of Chiropractic Examiners (NBCE) Part IV exam was originally developed to assure chiropractic licensing boards that applicants for licensure possessed the requisite skills and ability to perform these functions in a safe and effective manner, thereby protecting the public's health. Technological advances in DIM (primarily the transition from film-based to digital images) have mandated significant changes in the methods of testing current examinees.

In data collected in 2014 for the NBCE's 2015 practice analysis, 28.1% of chiropractors who took x-ray images in their offices used digital equipment to obtain images of their patients.[1] Since the radiography industry was rapidly undergoing technological change, chiropractors without radiographic equipment frequently referred their patients to imaging facilities with digital equipment and then reviewed the resultant digital images. Considering these developments, the NBCE directed its staff to redevelop the DIM component of the Part IV exam to use digital images.

In 2016 and 2017, the NBCE pilot tested a modified digital version of the DIM exam at 5 chiropractic colleges with promising results.[2,3] The 2018 Part IV Test Committee then selected and approved digital images for the modified DIM component, which was administered at the subsequent Part IV examination in November 2018. The objective of this article is to introduce the changes in the exam and to evaluate the possible effects of these changes.

BACKGROUND

Digital Radiographs in Testing

In response to the technological advancements in health care, boards and organizations responsible for pre-licensure and certification testing started making headway

in adopting digital technology for testing purposes. To construct items with digital images, testing organizations had to change the platform of exam delivery—tests needed to be delivered on computers. The American Board of Radiology began to develop computer-based exams in 1997 when a flexible examination platform, adapted to the graphical needs of an image-based item, was developed.[4] Since then, the medical specialty boards of pathology, pediatrics, family practice, internal medicine, neurology, and obstetrics and gynecology have all moved their exams to computer-based testing.[5] Today, the fact that the future belongs to digital radiology is well recognized by all areas of health care. The transition to digital imaging in DIM is supported by the industry,[6,7] and best practices in digital radiography have been developed and followed.[8]

The NBCE Practice Analysis survey[1] inquired whether students in chiropractic training programs have access to digital x-ray imaging. The responses were 73.3% "yes" and 26.7% "no" in 2008 and 100% "yes" in 2014. In 2014, the NBCE surveyed the radiology faculty in chiropractic institutions to determine the extent of usage of digital imaging in chiropractic colleges. Sixty-nine percent of the respondents indicated that digital images were used in patient clinics in 100% of the cases, 23% indicated that digital images were used in 75% of the cases, and 8% indicated that digital images were used in 0% of the cases.

With evidence of increased implementation of digital radiography in both chiropractic practice and chiropractic education, the NBCE made the decision to move away from the use of conventional radiographic images in the Part IV exam. In early 2015, the NBCE began a feasibility study for the transition to digital imaging, looking at various modes of delivery, methods of building an image library, and identifying chiropractic campuses to begin pilot studies. Five chiropractic educational institutions were identified for the pilot examination, and in the period between July 2016 and June 2017, a 10- and 20-station/image exam form was administered. Following the pilot exams, a 2-way univariate analysis of variance was conducted to investigate the effects of test form and test site (institutes) on the test scores. The analysis performed showed that when test forms and test sites are considered, the variability in scores could be explained only by differences in performance among various test sites while controlling for the effect of test form and the interaction effect.[2] Next, the challenge was to decide which psychometric models to use for scoring the newly refined exam.

Models for Polytomously Scored Items

Each of the DIM digital presentations asks for 2 interpretation responses from the examinee. Since these 2 questions are linked to the same stimulus (the station's digital images), they cannot be treated as independent items. A wide range of item response theory (IRT) models have been proposed to handle polytomous items with possible local dependency when testlets are formed, such as the Rating Score Model,[9] Partial Credit Model (PCM[10]), Generalized Partial Credit Model (GPCM[11]), and Graded Response Model (GRM[12]).

Samejima[12] developed a logistic model for graded responses in which the probability that a student i with a particular ability level θ_i will provide a response to an item j of the category k is the difference between the cumulative probability of a response to that category or higher and the cumulative probability of a response to the next highest category or higher. Let's consider the following:

$$P_{ijk}(\theta) = P'_{ijk}(\theta) - P'_{ijk+1}(\theta),$$

$$P'_{ijk}(\theta) = \frac{1}{1 + \exp[-Da_j(\theta - b_{jk})]},$$

where b_{jk} is the difficulty parameter for category k_j, a_j is the discrimination parameter for item j, and D is the scaling constant 1.702.[12,13] Samejima's GRM is classified as a difference model because the category probability is calculated by the difference of 2 cumulative probabilities.[10]

Another model for ordered-categorical responses was developed by Masters.[10] In this PCM, the probability that a student i will provide a response x on item j with M_j thresholds is a function of the student's ability and the difficulties from the M_j thresholds in item j is given by the following:

$$P_{ijx} = \frac{exp\sum_{x=0}^{X}(\theta_i - b_{jx})}{\sum_{m=0}^{M}\left(exp\sum_{x=0}^{m}(\theta_i - b_{jx})\right)},$$

where $x = 1,2,\ldots, M_j$ is the count of successfully completed thresholds and

$$\sum_{x=0}^{0}(\theta_i - b_{jx}) == 0.$$

In the PCM, the b_{jk} is referred to as a step difficulty parameter with a category x, and the higher value of b_{jk}, the more difficult a category is relative to other categories within an item. The b_{jk} can also be interpreted as the intersection point of 2 adjacent category response curves.[14] The PCM is a divide-by-total model, as it is written as an exponential divided by the total of exponentials.[15]

Muraki[11] developed a generalization of the PCM that allows the items within an instrument to have a different discrimination parameter as in the case of the GRM[12]; this is referred to as the GPCM. The GPCM substitutes a discriminant parameter a_j into Masters's PCM:

$$P_{ijx} = \frac{exp\sum_{x=0}^{X}a_j(\theta_i - b_{jx})}{\sum_{m=0}^{M}\left(exp\sum_{x=0}^{m}a_j(\theta_i - b_{jx})\right)},$$

where

$$\sum_{x=0}^{0}a_j(\theta_i - b_{jx}) = 0.$$

Muraki[11] described the discriminant parameter a_j as "the degree to which categorical responses vary among items as θ level changes."

Ostini and Nering[16] concluded that the theoretical choice between the GPCM and GRM is somewhat arbitrary and that the difference between the 2 models is purely mathematical. They suggest that the best method to select a model should begin with a consideration of the data characteristics.

In the area of cognitive testing, Cook et al[17] systematically compared the GPCM and the GRM in the context of testlet scoring for the fall 1994 administration of the Scholastic Assessment Test I. They found that the correlation between theta estimates was very high (r = .987) and that both models exhibited a good model fit across examinees' ability ranges as assessed by the plots of empirical and theoretical probabilities. The only analysis that produced notable differences between the 2 models was that the GRM had greater information function than the GPCM across most of the examinees' ability ranges. However, they showed that the greater information obtained from the GRM was explained partly by the higher value of the discrimination parameter estimates by the model.

Naumenko[18] compared the GRM and GPCM for testlet scores with a task-based simulation certified public accountant exam. The results showed the close relationships between the GRM and GPCM ability estimates. The examination of item fit statistics revealed a basically equivalent fit of both models to the exam data. When comparing the information functions, the GRM provided greater information over a wider range of ability estimates than the GPCM. The author noted that the comparison of information function between the 2 different models could be theoretically challenging because the calculation of discrimination parameters differed for the different models.

In the area of noncognitive tests, Baker et al[13] compared the performance of the Samejima GRM with Masters's PCM in a questionnaire about subjective well-being with a 5-point Likert scale. They found that Samejima's GRM outperformed Masters's PCM model by being more robust to violations of the unidimensionality assumption and better fitting the data. Further, they noted that Masters's[10] PCM may be more applicable in situations where the items meet the assumption of an equal slope parameter across items.[13]

Maydeu-Olivares[19] compared the fit for both the GRM and the GPCM to the Social Problem Solving Inventory with a 5-point Likert scale. The results showed that the GRM consistently outperforms all divide-by-total models, including the GPCM, having the smallest mean of the χ^2/df ratio for item pairs and triplets for all scales.

Research Questions

Given the above, the following research questions were addressed by this study:

- RQ1: Are the classical test theory (CTT)–based parameter estimates comparable across test forms and across May and November DIM administrations?
- RQ2: Are the IRT-based parameters comparable across test forms and across May and November DIM testing administrations?
- RQ3: Is there evidence of decision consistency/accuracy across the May and November DIM testing administrations?

METHODS

Study Objectives

The goals of this research were to introduce the changes made in the DIM portion of NBCE's Part IV exam and to study the effects of these changes on the test items and on examinees' performance. We employed methodologies based on CTT and IRT.[20,21] We also studied the decision accuracy of the redesigned DIM exam. One of the changes made to the exam was to substitute x-ray films on view boxes with digital images. Therefore, it is natural to expect that this change may account for some variability in item parameters and/or in test scores. Nevertheless, this was not an investigation of the effect of administration mode. Based on our pilot test data, we assumed that if such an effect existed, it would have minimal impact on the item difficulty parameters and even less on test scores.

Participants

This study used data from operational administrations of the NBCE's Part IV exams; therefore, all participants were chiropractic students who are within 6 months of the graduation or graduates of an eligible chiropractic college and who have passed the Part I exam. The study was approved by the institutional review board of the NBCE. The number of test takers in May 2018 was $n = 1424$; of them, $n = 1152$ were nonaccommodated, first-time examinees (norming group). There were $n = 796$ examinees who were administered Form 1 of the exam and $n = 628$ who were administered Form 2. There were $n = 1298$ examinees who attempted the redesigned Part IV exam in November 2018; of them $n = 1169$ were of the norming group. The number of examinees who were administered Form 1 in November was $n = 867$, while $n = 431$ were administered Form 2.

The average DIM raw scores in May were mean (M) = 28.13, SD = 4.86, for Form 1 and M = 27.78, SD = 4.94, for Form 2. The average DIM raw scores for November were M = 29.95, SD = 4.17, for Form 1 and M = 28.31, SD = 4.61, for Form 2.

Measures

In May 2018, the DIM portion of the exam consisted of 10 stations with 2 items per station. Each item had 10 response choices, and 2 correct responses were required to obtain full credit. The examination in May was administered on film (on lighted view boxes). In November, the DIM portion consisted of 20 stations with 2 items per station. Each item had 4 response choices, and the correct response was required to obtain full credit. The exam in November used digital images displayed on computer monitors.

The time allowed per station was 4 minutes in May and 2 minutes in November. In May, the first item in each station inquired about the x-ray findings present on the

film, while the second item was developed to address the impression/diagnosis, case management, or sequela of the condition. In November, the first item probed the impression/diagnosis, and the second item addressed case management or sequela.

Scoring Rubrics

The test items in May were scored polytomously from 0 to 4. The following algorithm was employed to convert raw to scored responses in May: 0 out of 4 results in a score of 0, 1 out of 4 results in 1, 2 out of 4 results in 2, 3 out of 4 results in 3, and 4 out of 4 results in 4. The test items in November were scored polytomously from 0 to 2. The scoring rubric implemented in November was as follows: 0 out of 2 resulted in a score of 0, 1 out of 2 resulted in 1, and 2 out of 2 resulted in 2.

Classical Item Analysis

After receiving all examinee response data, implementing scoring rules, validating the responses in data files, and applying agreed-on valid case criteria to the data, classical item analysis to evaluate item difficulties and item discriminations was performed. The classical item analysis was conducted on the operational and field test items to collect information about item performance. The analysis is called "classical" because all statistical estimates calculated during this procedure are based on CTT. The basic assumption of CTT is that the overall (observed) score has 2 components—the true score, which represents the actual ability of a test taker,[22] and an error component, which by itself is a combination of systematic and random errors.[23–26] CTT assumes that the true score is relatively stable, while the random error, on the other hand, is not. Therefore, supposing that all systematic variability is accounted for by true score(s), an average of a measurement that was taken a reasonable number of times will approximate the true score.

Classical item analysis provides information about 2 concepts of interest to test developers and psychometricians—the estimates of item difficulty and the estimates of item discrimination. Item difficulty is a relative concept, as the same item may be difficult for 1 population while easy for another. Psychometricians make difficulty judgments by taking into account the content on the test, the purpose of the test, and the population to which the test is to be administrated. Thus, for a dichotomously scored test item, the estimate of item difficulty (p value) in the framework of CTT is computed by calculating the proportion of examinees who answered that item correctly. Items with difficulty values closer to 1.0 are considered easier items, while items with difficulty estimates closer to 0 are considered harder. Desired p values generally fall within the range of .25–.95.

For polytomous items, the difficulty is measured by taking the mean of the item score. The average item score could range from 0 to the maximum possible score points for the item. To help with the interpretation, the item average is usually expressed as a percentage of the maximum possible score, which is equivalent to the p value in dichotomous items. Numerous factors may cause items to fall outside the desired difficulty ranges. Items may not perform as expected due to the lack of familiarity with the content or the item type. Test administrators may wish to consider these items for future use based on the importance of the item content or the need to measure with more precision the performance of examinees with very high or very low ability levels.

The purpose of most tests is to classify the examinee population into mastery/nonmastery groups according to the construct measured by the test.[24] Therefore, it is important that a test item effectively discriminate between these 2 populations. The statistic that relates the performance on an item to the total score obtained on the test is called the item-total correlation. In the CTT framework, this statistic serves as the index of discrimination. For polytomously scored categorical items, the polyserial correlation is computed as an estimate of the relation between a continuous variable and an ordinal, categorical variable.[27]

Item-total correlation can range from −1.0 to +1.0. Desired values are positive and greater than .20. A negative item-total correlation indicates that low-ability examinees outperformed the high-ability examinees, which may suggest a range of problems, from a mis-key during the scoring process to serious problems with content development.

In this study, the classical item analysis was conducted in R[28] using the "psychometric" package.[29] Frequency distributions were constructed to identify items with few or no observations at any score point. In addition, the average item scores and the estimates of the correlation with criterion (total score) were calculated. The correlations were derived without excluding the item under consideration from the total score.

Overview of Statistical Modeling

The choice of statistical model used to fit data depends on the type of the test item, the design of the response options, and the rubrics used to convert raw scored responses. For items scored dichotomously, when each answer is scored as correct or incorrect, a plethora of logistic IRT models is available.[20,21] However, when the response is scored on the ordinal scale,[30] more universal models are available for use.[31] Various polytomous IRT models had been developed to fit ordered categorical responses. These models could be classified into 3 categories: adjacent category models, or generalized partial credit models; cumulative probability models, or graded response models; and sequential models, also known as continuation ratio models. Typically, in these models, higher levels of responses are associated with higher student ability levels.

Polytomous IRT models were developed to describe the probability that a response falls into a particular category conditional on an examinee's ability level and item parameters.[32] Similar to the framework of generalized linear models, where models for ordinal response data are extended from generalized linear models for binary response,[33] most of the IRT models for polytomous data are extended from IRT models for binary responses. The

item characteristic curve, which is a function that connects test taker's ability to the probability of a response, is usually constructed only for the correct response in IRT models for binary data. This is because the probability of the incorrect response is simply 1: the probability of the correct response.[31] This is not the case for polytomous models— the curve is constructed for each scored response. For example, an item scored 0 through 3 will have 4 curves corresponding to the probability of each response while conditional on the ability level.

IRT Calibration

The term calibration in psychometric context corresponds to the fit of statistical models to the item response data. The purpose of IRT calibration and scaling is to place all operational and field test items onto a common scale. There were 2 operational DIM forms used in May and 2 in November. The calibration was performed within forms for the May and the November administrations.

IRT is a collection of measurement models that connect a test taker's score on the latent trait to the probability of correct response based on the observed item responses.[31] Traditionally, IRT models assume unidimensionality and local independency.[32] Under the assumption of unidimensionality, a single latent trait should account for all systematic variability among item responses.[34] The assumption of local independence states that the probability of a correct response on 1 item is not influenced by the performance on any of the other items.[35] We tested the assumption of unidimensionality with factor analysis. To diagnose the assumption of local independence, Q3 statistics were estimated for Rasch models in each form,[36] resulting in estimates for every item pair, which are the correlations between the item residuals after fitting the Rasch model. Yen[37] provides guidance for the assessment of local independence: "The expected value of Q3, when local independence holds, is approximately $-1/(n-1)$." There was no evidence of violation of local independence assumptions among the DIM items.

Graded Response Model

The GRM extends from the logistic positive exponent family of models for dichotomous responses.[12] The GRM calculates the probabilities based on the 2PL model specification while estimating difficulty parameters for every step of the item and a single discrimination parameter.[18]

Baker et al[13] compared the performance of Samejima's[12] logistic model with Masters's[10] model using psychological data with multiple response categories. They found that Samejima's logistic model outperformed Masters's model by being more robust to violations of the unidimensionality assumption and better fitting the data. Further, they noted that Masters's PCM may be more applicable in situations where the items meet the assumption of equal slope parameter across items. The following section describes the GRM in statistical terms.

Let's consider an ordinal item j with k response categories; then let θ be the latent variable representing the latent trait being measured. Then the GRM specifies the trace function for the item:

$$P_j(\text{category}k|\ \theta) = P_j^*(k|\theta) - P_j^*(k+1|\theta),$$

where $P_j(\text{category } k|\theta)$ is the probability of a response in category k given that the latent trait is $P_j^*(k|\theta)$—the probability of the observed response in category k or higher minus $P_j^*(k+1|\theta)$—the probability of the observed response in category higher than k. If $k = 1, 2, \ldots, K-1$, then the probabilities of responses may be generalized in the following way:

$$P_j^*(k|\theta) = \begin{cases} 1 & \text{if } k = 0, \\ 0 & \text{if } k = K, \\ \dfrac{1}{1+\exp\Big(-\alpha_j(\theta_i - \beta_{jk})\Big)} & \text{otherwise,} \end{cases}$$

where α_j is the discrimination parameter for item j and β_{jk} is the difficulty parameter for the response in category k within item j. That is, if an item is scored 0, 1, 2, 3, then, given that a response is provided, the probability that the response is in category 0 or higher is unity. The probability that the response is in category 4 is 0. Finally, the probability that the response is somewhere between 0 and 3 is given by[12,38]

$$\frac{1}{1+\exp\Big(-\alpha_j(\theta_i - \beta_{jk})\Big)}.$$

Decision Consistency

We took the slant of reliability for the decision consistency study. The synonyms for reliability are dependability, stability, consistency, reproducibility, predictability, and lack of distraction. One of the approaches to establish reliability is to inquire how much error of measurement is in a measuring instrument.[26] Since the object of our attention in this study was a classification instrument (a test), we believed it would be appropriate to measure the extent of error made in classification. Furthermore, since the test is scored using IRT, we decided to investigate the reliability of classification into an ability category.

According to Baker,[39] each examinee responding to a test item possesses some amount of the underlying ability. Thus, one can consider each examinee to have a numerical value, a score that places the examinee on the ability scale. The basic premise of IRT is that the probability of correct response on an item is a function of ability, denoted by θ. The ability estimates are derived during the process of test calibration. Next, the examinees are categorized according to their estimated ability based on their item responses. In this study, we compared the frequency distributions of examinees in different ability categories.

To ensure longitudinal comparison of scores, decision consistency is required when changes are made to an assessment instrument. However, the desired levels of decision consistency do not guarantee that testing results necessarily reflect examinees' true ability.[40] While decision consistency suggests reliability, the idea of the instrument testing what it is supposed to test is the most critical aspect

of validity.[23] This article does not attempt to make a content validity study; therefore, our results should be interpreted with caution.

RESULTS

Overview

The results were analyzed using rigorous, widely accepted statistical methodologies for evaluating validity and reliability; the statistical fit of the models used for item calibration, equating, and scaling; and longitudinal consistency of the exam. Classical statistics were used for the assessment of item difficulty and the relation to the total test score. Item difficulties along with item discrimination were calculated using IRT.

Data Cleaning and Classical Item Analysis

The classical item analysis was conducted in R using the "psychometric" package.[29] In preparation for item analysis, all item responses were reviewed to verify that the data were free of errors. We looked for data entry errors, data merge errors, missing values, and out-of-range responses. The purpose of data cleaning procedures was to ensure that the psychometric analyses were conducted on a valid set of examinee responses.

Tables 1 through 4 present the results of item analysis for DIM administrated in May (Tables 1 and 2) and in November (Tables 3 and 4). The statistics used in this step, *p* values, polyserial correlations, and reliability coefficients were derived from CTT. Item analyses were conducted by form and by administration. The primary purpose of these analyses was to evaluate the quality of a test item; therefore, in addition to operational items, field-test items were included.

The stations on Form 1 administered in May showed consistent performance in terms of difficulty and discrimination. The difficulty estimates ranged from M = 1.9, SD = 1.34, to M = 3.69, SD = .61. The scoring rubric for DIM in May ranged from 0 to 4. Lower numbers correspond to more difficult items, whereas higher numbers indicate easiness. The average form difficulty was M = 2.85, SE = .6. The average correlation with the criterion was $\bar{r} = .42$, which is within the acceptable range. The item-total correlations can range from −1.0 to +1.0. Desired values are positive and larger than .20. There were no stations with extremely low or negative correlations on Form 1 administered in May.

The difficulty and discrimination parameter estimates on Form 2 administered in May were within expected ranges. The difficulties ranged from M = 1.87, SD = 1.23, to M = 3.66, SD = .7. For the 10 stations on the form, the average difficulty was M = 2.81, SE = .6. The average correlation with criterion was $\bar{r} = .45$. The 2 DIM forms administrated in May reveal consistency in terms overall difficulty and discrimination calculated using classical methodology.

The DIM exam in November was scored using a rubric ranging from 0 to 2. The average difficulty on Form 1 administered in November was M = 1.51, SE = .33. The average discrimination (correlation with the criterion) was $\bar{r} = .3$. The average difficulty for Form 2 administered in November was M = 1.43, SE = .33. The average discrimination was $\bar{r} = .31$. Similar to the forms administered in May, the November forms revealed consistency in difficulty and discrimination.

Table 1 - Descriptive Statistics Based on Classical Test Theory, May 2018 Administration, Form 1

Station	Response	*N*	Percent	Mean (SD)	*r* w/Crit.
1	0	0	0%	3.69 (.61)	0.25
	1	6	1%		
	2	33	5%		
	3	113	17%		
	4	502	77%		
2	0	33	5%	2.26 (1.11)	0.42
	1	145	22%		
	2	190	29%		
	3	192	29%		
	4	94	14%		
3	0	11	2%	3.39 (1.12)	0.45
	1	77	12%		
	2	37	6%		
	3	50	8%		
	4	479	73%		
4	0	109	17%	1.9 (1.34)	0.44
	1	193	30%		
	2	107	16%		
	3	146	22%		
	4	99	15%		
5	0	32	5%	2.99 (1.29)	0.52
	1	98	15%		
	2	60	9%		
	3	116	18%		
	4	348	53%		
6	0	37	6%	3.39 (1.12)	0.37
	1	30	5%		
	2	19	3%		
	3	126	19%		
	4	442	68%		
7	0	56	9%	2.41 (1.31)	0.5
	1	126	19%		
	2	153	23%		
	3	133	20%		
	4	186	28%		
8	0	15	2%	3.19 (1.01)	0.46
	1	43	7%		
	2	62	9%		
	3	215	33%		
	4	319	49%		
9	0	11	2%	2.3 (.83)	0.32
	1	86	13%		
	2	294	45%		
	3	225	34%		
	4	38	6%		
10	0	31	5%	2.98 (1.21)	0.44
	1	69	11%		
	2	93	14%		
	3	152	23%		
	4	309	47%		

r w/Crit., correlation with the criterion.

Table 2 - Descriptive Statistics Based on Classical Test Theory, May 2018 Administration, Form 2

Station	Response	*N*	Percent	Mean (SD)	*r* w/Crit.
1	0	65	13%	1.87 (1.23)	0.47
	1	155	31%		
	2	117	23%		
	3	101	20%		
	4	60	12%		
2	0	25	5%	2.45 (1.26)	0.48
	1	129	26%		
	2	73	15%		
	3	141	28%		
	4	130	26%		
3	0	5	1%	3.27 (.99)	0.46
	1	36	7%		
	2	59	12%		
	3	119	24%		
	4	279	56%		
4	0	6	1%	3.12 (1.09)	0.38
	1	36	7%		
	2	126	25%		
	3	53	11%		
	4	277	56%		
5	0	8	2%	3.19 (1.02)	0.48
	1	42	8%		
	2	44	9%		
	3	157	32%		
	4	247	50%		
6	0	45	9%	2.32 (1.27)	0.48
	1	111	22%		
	2	80	16%		
	3	164	33%		
	4	98	20%		
7	0	54	11%	2.05 (1.13)	0.43
	1	97	19%		
	2	160	32%		
	3	142	29%		
	4	45	9%		
8	0	14	3%	3.23 (1.12)	0.46
	1	51	10%		
	2	28	6%		
	3	117	23%		
	4	288	58%		
9	0	1	0%	3.66 (.7)	0.37
	1	13	3%		
	2	20	4%		
	3	88	18%		
	4	376	76%		
10	0	5	1%	2.93 (.98)	0.47
	1	39	8%		
	2	109	22%		
	3	178	36%		
	4	167	34%		

r w/Crit., correlation with the criterion.

Calibration

The purpose of IRT calibration and scaling is to place the scores of different testing administrations on a common difficulty scale. For calibration of DIM items, the GRM[12] was employed. The operational calibration was performed using IRTPRO 4.2 for Windows[41]—a computer program that provides an ability to calibrate item responses using a plethora of IRT models for dichotomous and polytomous data. For this study, however, in addition to operational results, the calibration was replicated in R using the "ltm" package,[42] achieving a perfect match between both sets of results. The calibration was performed within form and within administration; therefore, 4 sets of results are presented: Form 1 and Form 2 administered in May 2018 and Form 1 and Form 2 administered in November 2018.

Table 5 presents the discrimination (a) and category difficulties (b's) along with standard errors associated with these estimates for Form 1 administered in May. The last column gives the $S - X^2$ generalized to polytomous models.[43] Two aspects of the results are notable. First, no items showed statistically significant misfit; however, several items had b_1 estimates outside of expected ranges, indicating that the category may be too easy for the population of examinees. The difficulty averages for the form were $\bar{b}_1 = -5.55$, $\bar{b}_2 = -2.85$, $\bar{b}_3 = -1.26$, and $\bar{b}_4 = .91$. Second, the standard error estimates were reasonable. Station 1 failed to attract responses in the fourth category, resulting in a missing estimate for b_4 and its associated standard error.

Table 6 shows calibration results for the Form 2 administered in May. All items but 1 (station 4) revealed good fit of the model. The fit index calculated for station 4 revealed statistical significance ($p < .01$), indicating a misfit. Furthermore, the difficulty estimate associated with the item was out of range ($b_1 = -10.65$); therefore, the item was removed from the second round of operational calibration and was not included in the calculation of the total score. The difficulty averages for Form 2 were consistent with the estimates calculated for Form 1: $\bar{b}_1 = -5.38$, $\bar{b}_2 = -2.58$, $\bar{b}_3 = -1.01$, and $\bar{b}_4 = .91$. The discrimination estimates for both forms were reasonable.

Table 7 gives estimates of the discrimination and difficulty for the 20 items (stations) on Form 1 administered in November. For the majority of the items on the form, the difficulty estimates were reasonable. Two stations revealed out-of-range estimates for b_1 (stations 4 and 9). These items were removed from the second round of operational calibration and calculation of the total score. Station 14 indicated a poor fit of the model; however, the difficulty estimates associated with the item were reasonable. Thus, the station was not removed from further consideration. The averages of the difficulty estimates on the form were $\bar{b}_1 = -5.52$ and $\bar{b}_2 = -2.71$.

Table 8 provides results for Form 2 administered in November. Analogously to Form 1, the estimates were reasonable for the majority of the items. While there were no items with poor statistical fit, stations 16 and 20 revealed out-of-range difficulty estimates. These stations were removed from the second round of operational calibration and calculation of the total score. The difficulty averages for Form 2 were $\bar{b}_1 = -5.77$, $\bar{b}_2 = -1.23$.

The calibration results showed consistency of the test before and after the change. The averages of the parameter

Table 3 - Descriptive Statistics Based on Classical Test Theory, November 2018 Administration, Form 1

Station	Response	*N*	Percent	Mean (SD)	*r* w/Crit.
1	0	95	12%	1.69 (.68)	0.33
	1	57	7%		
	2	633	81%		
2	0	28	4%	1.91 (.39)	0.16
	1	16	2%		
	2	741	94%		
3	0	78	10%	1.66 (.65)	0.33
	1	113	14%		
	2	594	76%		
4	0	6	1%	1.83 (.39)	0.19
	1	120	15%		
	2	661	84%		
5	0	301	38%	1.35 (.64)	0.35
	1	135	17%		
	2	349	44%		
6	0	30	4%	1.9 (.41)	0.21
	1	19	2%		
	2	736	94%		
7	0	96	12%	1.68 (.68)	0.26
	1	56	7%		
	2	633	81%		
8	0	93	12%	1.22 (.64)	0.25
	1	426	54%		
	2	266	34%		
9	0	124	16%	1.58 (.75)	0.18
	1	78	10%		
	2	583	74%		
10	0	456	58%	.73 (.91)	0.35
	1	83	11%		
	2	246	31%		
11	0	61	8%	1.82 (.56)	0.26
	1	14	2%		
	2	710	90%		
12	0	301	38%	1.06 (.91)	0.47
	1	135	17%		
	2	349	44%		
13	0	96	12%	1.69 (.68)	0.3
	1	51	6%		
	2	638	81%		
14	0	140	18%	1.2 (.72)	0.36
	1	352	45%		
	2	293	37%		
15	0	379	48%	1.02 (.99)	0.37
	1	15	2%		
	2	391	50%		
16	0	111	14%	1.54 (.73)	0.38
	1	138	18%		
	2	536	68%		
17	0	61	8%	1.83 (.55)	0.24
	1	14	2%		
	2	710	90%		
18	0	133	17%	1.43 (.76)	0.33
	1	186	24%		
	2	466	59%		
19	0	119	15%	1.52 (.74)	0.37
	1	138	18%		
	2	528	67%		
20	0	113	14%	1.62 (.72)	0.25
	1	70	9%		
	2	602	77%		

r w/Crit., correlation with the criterion.

estimates are comparable across the forms and administrations. The fit of the IRT model to the data is acceptable, and the number of items excluded from operational procedures is minimal. Further, the marginal reliability estimates for thetas were high and ranged from .78 to .85 for May and from .82 to .88 in November.

Figures 1 through 4 present item characteristic curves for DIM test stations on both forms administered in May and November. The rubrics used for the forms administered in May had 5 response options, while rubrics used for forms administered in November had 3. Each plot represents an item (station); the x-axis is the ability scale (theta) depicted in standardized (*z* score) units. The y-axis is the probability of the response conditional on the ability level. Each curve in the plot represents a response option.

For polytomously scored items, we expect that the probability of the response associated with the full credit will proliferate with the increase in the ability level. With minor exceptions, (eg, item 9 on May's Form 1), all plots show a monotonic increase of the curve associated with full-credit response as a function of ability level.

Decision Consistency

The consistency of classification of members of the same group into the same category by a testing instrument is very important for longitudinal comparison of the scores. In criterion-referenced testing, a common approach to study decision consistency is relative to the cut scores—when the same examinees are being tested two or more times, they should be classified with the same consistency while controlling for type I and type II errors. However, the test takers in May and November were different; therefore, we took a different approach to consistency study. We created 12 ability intervals ranging of .5 from −3.0 to 3.0, and we compared the count of examinees classified in each ability interval between testing administrations.

Table 9 presents results of classification consistency for the norming group (nonaccommodated, first-time test takers) and the overall sample for May and November administrations. It is evident from the table and the plots (Fig. 5) that the distributions are normal with the majority of test takers being clustered around the mean as expected. The minor between-category deviations in the counts of examinees are due to the random error.

DISCUSSION

Technological advances in the field of medical imaging required educational programs and testing organizations

Table 4 - Descriptive Statistics Based on Classical Test Theory, November 2018 Administration, Form 2

Station	Response	*N*	Percent	Mean (SD)	*r* w/Crit.
1	0	41	11%	1.64 (.67)	0.31
	1	55	14%		
	2	288	75%		
2	0	112	29%	1.31 (.89)	0.37
	1	40	10%		
	2	232	60%		
3	0	42	11%	1.65 (.67)	0.33
	1	51	13%		
	2	291	76%		
4	0	36	9%	1.8 (.59)	0.39
	1	4	1%		
	2	344	90%		
5	0	71	18%	1.36 (.78)	0.39
	1	102	27%		
	2	211	55%		
6	0	39	10%	1.67 (.65)	0.43
	1	50	13%		
	2	295	77%		
7	0	126	33%	1.03 (.83)	0.3
	1	119	31%		
	2	139	36%		
8	0	45	12%	1.62 (.69)	0.36
	1	57	15%		
	2	282	73%		
9	0	43	11%	1.56 (.73)	0.38
	1	19	5%		
	2	322	84%		
10	0	54	14%	1.56 (.73)	0.38
	1	62	16%		
	2	268	70%		
11	0	149	39%	1.16 (.96)	0.26
	1	23	6%		
	2	212	55%		
12	0	135	35%	1.22 (.94)	0.26
	1	31	8%		
	2	218	57%		
13	0	22	6%	1.39 (.59)	0.29
	1	190	49%		
	2	172	45%		
14	0	68	18%	1.64 (.77)	0.33
	1	4	1%		
	2	312	81%		
15	0	221	58%	.82 (.97)	0.37
	1	13	3%		
	2	150	39%		
16	0	12	3%	1.3 (.54)	0.14
	1	219	57%		
	2	153	40%		
17	0	39	10%	1.44 (.67)	0.28
	1	136	35%		
	2	209	54%		
18	0	108	28%	1.41 (.9)	0.22
	1	9	2%		
	2	267	70%		
19	0	162	42%	.98 (.91)	0.32
	1	66	17%		
	2	156	41%		
20	0	1	0%	1.82 (.39)	0.18
	1	67	17%		
	2	316	82%		

r w/Crit., correlation with the criterion.

in all health care fields to reexamine their curricula and testing content. A 2014 survey of the chiropractic colleges in the United States revealed that only 8% of the doctor of chiropractic programs did not use digital images in their outpatient clinics and 15% did not utilize digital radiographs for educational or examination purposes. Although a small percentage of doctor of chiropractic programs were not utilizing digital images, private practice usage of digital imaging increased from 11.6% in 2009 to 28.1% in 2014.[1] With the increase in private practice use along with the movement of all chiropractic educational institutions in the direction of digital radiography, the NBCE began the transition of the DIM component of the Part IV exam from plain film to digital radiography.

Following the pilot examination, it was determined that a 20 station/image was preferable in order to decrease item bias and content underrepresentation. It was also determined that in order to standardize the examination experience, the NBCE would provide identical monitors to the college campuses for the inaugural digital DIM examination that occurred in November 2018. While in the pilot planning stage, the NBCE also initiated contact with practicing chiropractors to begin building a substantial digital library, and this continues today.

Limitations

The findings of this research were derived from examining only 2 test administrations—May 2018, prior to the change in DIM, and November 2018, after the change. The Part IV exam started the transition to IRT scoring in 2017 with May 2018 being the first administration when IRT scoring was fully implemented. Therefore, it is reasonable to assume that May and November test takers may not be representative of all examinees who take the Part IV exam. Further, due to the nature of the chiropractic profession and chiropractic education, the cohorts of examinees are not very large, yet the statistical models used in this study are large-sample techniques. While the stability of the model estimates was reasonable, larger sample sizes may have been beneficial.

CONCLUSION

Aside from introducing chiropractic practitioners, chiropractic faculty, and chiropractic students to the changes in the DIM component of the Part IV exam, the goal of this study was to examine the performance of the revised exam. From a psychometric perspective, we examined the performance of the items on the exam as

Table 5 - Graded Response Model Parameter Estimates for May Administration, Form 1, logit $a(\theta - b)$

Station	a	SE	b_1	SE	b_2	SE	b_3	SE	b_4	SE	χ^2 (df)
1	0.52	0.14	−9.53	2.57	−5.48	1.39	−2.41	0.6	N/A	N/A	45.01 (38)
2	0.55	0.11	−5.59	1.08	−1.93	0.38	0.47	0.17	3.45	0.66	85.18 (70)
3	0.94	0.17	−4.77	0.79	−2.28	0.34	−1.78	0.27	−1.26	0.19	58.49 (55)
4	0.49	0.11	−3.45	0.74	−0.35	0.18	1.08	0.28	3.69	0.79	85.26 (73)
5	0.92	0.15	−3.6	0.51	−1.76	0.24	−1.15	0.17	−0.18	0.1	61.37 (64)
6	0.5	0.13	−5.84	1.42	−4.53	1.09	−3.95	0.94	−1.55	0.39	57.89 (61)
7	0.75	0.12	−3.46	0.53	−1.45	0.23	0.04	0.11	1.35	0.23	71.0 (71)
8	0.89	0.14	−4.65	0.7	−2.98	0.42	−1.95	0.27	0.06	0.1	75.69 (61)
9	0.44	0.11	−9.51	2.31	−4.8	1.12	−1.43	0.37	2.47	0.6	60.32 (62)
10	0.62	0.12	−5.11	0.92	−2.96	0.52	−1.53	0.29	0.18	0.14	77.11 (69)

Table 6 - Graded Response Model Parameter Estimates for May Administration, Form 2, logit $a(\theta - b)$

Station	a	SE	b_1	SE	b_2	SE	b_3	SE	b_4	SE	χ^2 (df)
1	0.55	0.12	−3.66	0.75	−0.44	0.19	1.44	0.33	3.83	0.79	74.39 (70)
2	0.68	0.12	−4.64	0.82	−1.34	0.25	−0.33	0.15	1.67	0.31	96.61 (68)
3	0.86	0.15	−5.76	1.03	−3.15	0.5	−1.87	0.29	−0.34	0.12	61.45 (53)
4	0.42	0.12	−10.65	3.12	−5.8	1.64	−1.66	0.5	−0.57	0.26	101.06 (56)[a]
5	0.95	0.15	−4.8	0.75	−2.67	0.38	−1.8	0.26	0.02	0.11	49.06 (56)
6	0.62	0.12	−3.97	0.73	−1.41	0.28	−0.23	0.16	2.4	0.45	65.61 (70)
7	0.8	0.13	−2.95	0.45	−1.22	0.21	0.66	0.16	3.16	0.48	81.61 (66)
8	0.91	0.16	−4.31	0.69	−2.39	0.36	−1.88	0.28	−0.43	0.12	70.97 (56)
9	0.88	0.18	−7.48	1.77	−4.42	0.81	−3.36	0.59	−1.5	0.26	58.03 (38)
10	0.89	0.14	−5.56	0.93	−2.96	0.43	−1.09	0.18	0.88	0.17	56.64 (56)

[a] $p < .01$.

Table 7 - Graded Response Model Parameter Estimates for November Administration, Form 1, logit $a(\theta - b)$

Station	a	SE	b_1	SE	b_2	SE	χ^2 (df)
1	0.66	0.14	−3.26	0.63	−2.37	0.45	40.48 (34)
2	0.56	0.22	−6.14	2.25	−5.28	1.91	22.55 (23)
3	0.63	0.13	−3.74	0.71	−1.96	0.37	36.81 (35)
4	0.34	0.14	−14.64	6.06	−5.04	2.01	31.77 (22)
5	0.58	0.11	−4.13	0.76	0.46	0.15	40.71 (35)
6	0.89	0.23	−4.04	0.88	−3.43	0.73	29.35 (25)
7	0.45	0.13	−4.53	1.24	−3.3	0.9	31.21 (34)
8	0.29	0.1	−7.01	2.35	2.35	0.81	27.92 (37)
9	0.05	0.11	−32.69	70.74	−20.61	44.59	31.44 (34)
10	0.44	0.11	0.77	0.25	1.88	0.47	39.41 (33)
11	0.58	0.18	−4.33	1.19	−4.03	1.1	19.96 (23)
12	0.92	0.14	−0.61	0.11	0.28	0.1	29.56 (35)
13	0.59	0.14	−3.53	0.76	−2.64	0.57	21.80 (34)
14	0.58	0.11	−2.8	0.5	0.98	0.21	59.31 (37)[a]
15	0.46	0.11	−0.16	0.16	0.01	0.16	25.46 (29)
16	0.8	0.13	−2.51	0.37	−1.08	0.17	32.73 (35)
17	0.56	0.18	−4.64	1.34	−4.23	1.21	27.21 (26)
18	0.51	0.11	−3.31	0.68	−0.82	0.21	40.41 (36)
19	0.57	0.12	−3.18	0.63	−1.33	0.28	43.01 (36)
20	0.3	0.12	−5.97	2.29	−4.0	1.53	33.91 (33)

[a] $p < .01$.

Table 8 - Graded Response Model Parameter Estimates for November Administration, Form 2, logit $a(\theta - b)$

Station	a	SE	b_1	SE	b_2	SE	χ^2 (df)
1	0.57	0.18	−3.94	1.14	−2.06	0.6	30.80 (30)
2	0.45	0.15	−2.08	0.7	−1	0.39	33.68 (31)
3	0.7	0.19	−3.26	0.8	−1.8	0.44	29.62 (31)
4	0.88	0.28	−2.91	0.75	−2.77	0.71	26.36 (17)
5	0.8	0.17	−2.09	0.4	−0.29	0.15	39.48 (31)
6	1.37	0.29	−2.07	0.31	−1.18	0.18	20.52 (27)
7	0.27	0.13	−2.65	1.29	2.1	1.05	42.79 (34)
8	0.79	0.19	−2.83	0.6	−1.44	0.31	25.38 (31)
9	0.77	0.22	−2.97	0.75	−2.38	0.59	37.97 (26)
10	0.85	0.19	−2.41	0.47	−1.14	0.24	34.74 (29)
11	0.17	0.14	−2.79	2.35	−1.29	1.2	45.47 (32)
12	0.23	0.14	−2.78	1.71	−1.26	0.86	47.49 (31)
13	0.48	0.15	−6.03	1.81	0.46	0.26	34.02 (32)
14	0.55	0.19	−2.98	0.97	−2.85	0.93	22.68 (20)
15	0.62	0.17	0.53	0.22	0.77	0.26	32.09 (25)
16	0.08	0.14	−40.44	65.3	4.87	7.94	26.54 (26)
17	0.35	0.14	−6.32	2.52	−0.5	0.35	43.34 (34)
18	0.09	0.15	−10.77	18.37	−9.47	16.16	36.53 (24)
19	0.47	0.14	−0.72	0.3	0.85	0.34	35.36 (32)
20	0.38	0.19	−15.82	8.04	−4.16	1.95	27.24 (16)

well as the performance of the 2 cohorts—examinees who took the test in May and November 2018.

The diagnostic item analysis revealed similarity in item functioning across test forms and across administrations. The distributions of responses across response options were reasonable, and the correlations with the criterion were in the expected ranges. The IRT models displayed reasonable fit to the data. There were only 2 occasions where fit indices showed a misfit at the item level—these were eliminated from the scoring process. The averages of

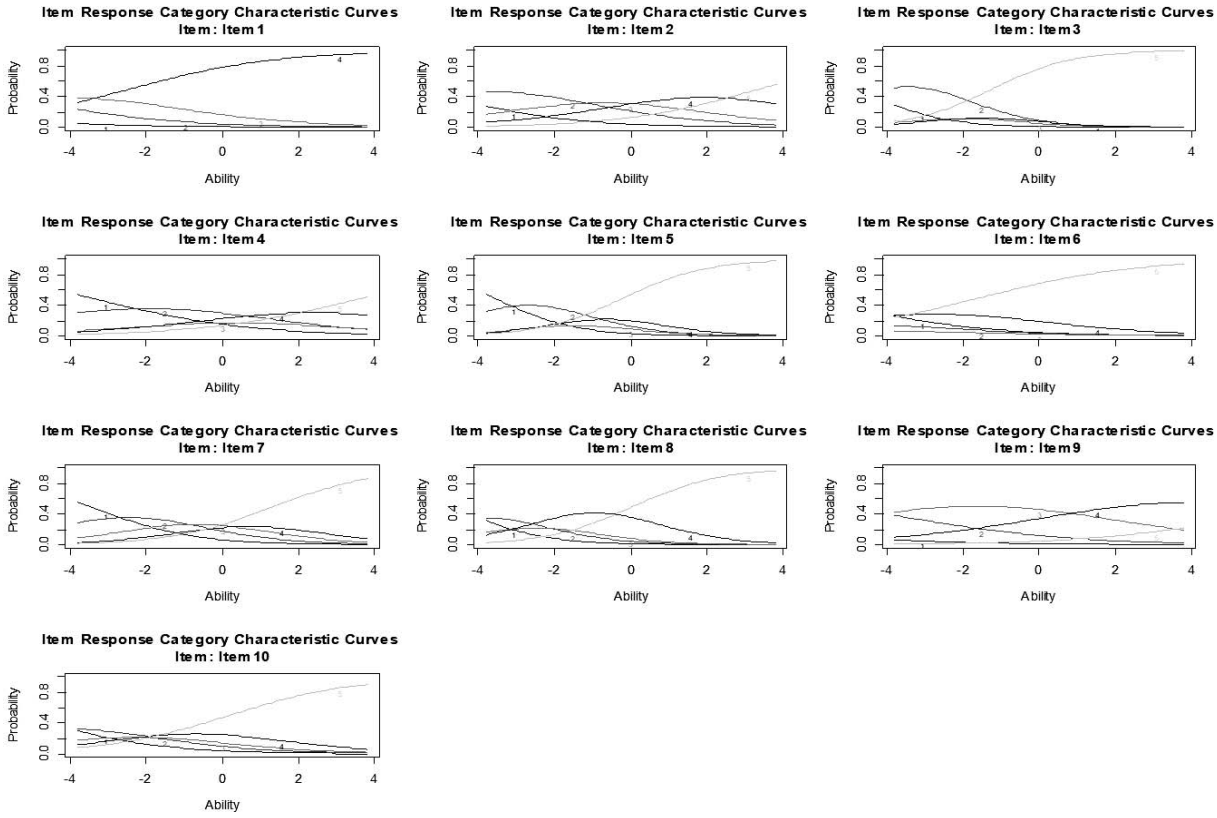

Figure 1 - Item characteristic curves for Form 1, administered in May 2018.

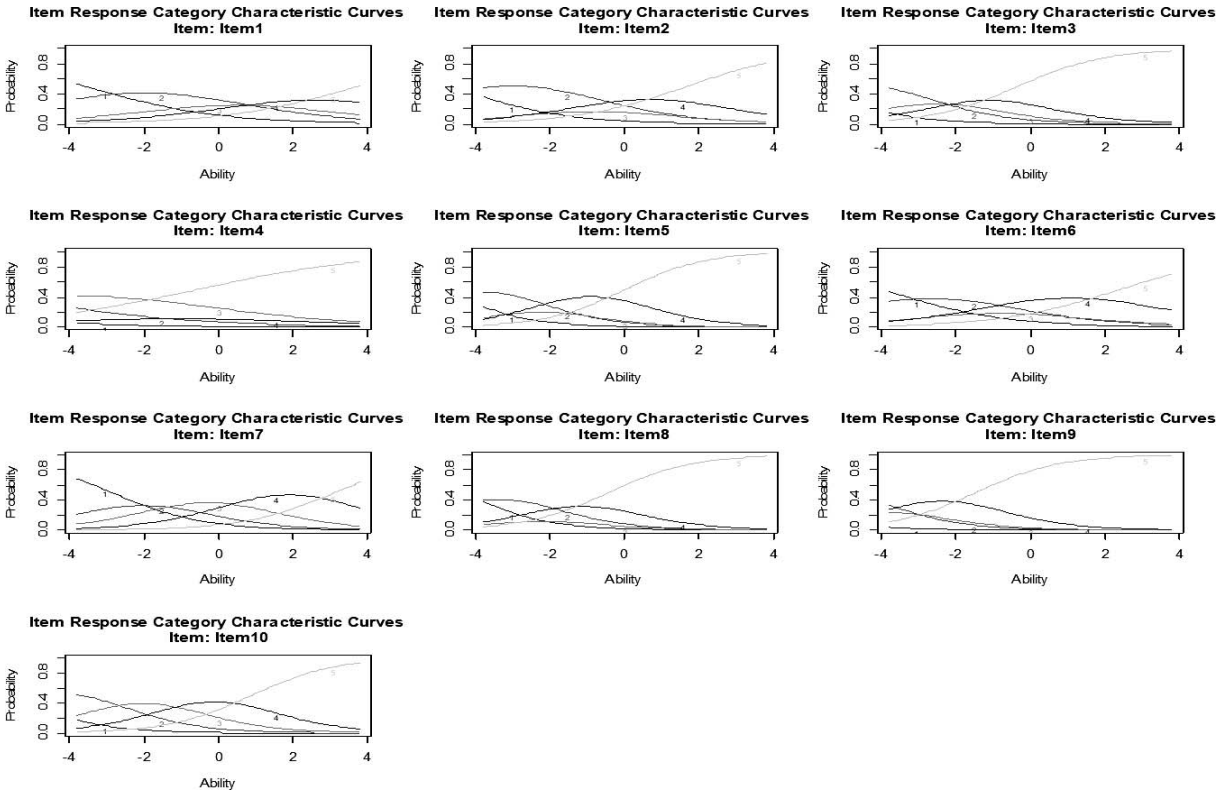

Figure 2 - Item characteristic curves for Form 2 administered in May 2018.

the IRT parameters were similar across test forms and across administrations. The classification of test takers into ability (theta) categories was consistent across norming/all groups, across test forms, and across administrations.

The transition to digital DIM was well thought through, relied on empirical findings, and was aligned with the current chiropractic curricula in the United States. This research signifies a first step in the evaluation of the transition to digital DIM. We hope that this study will spur further DIM research. In addition, we hope that the results prove to be useful for chiropractic faculty, chiropractic students, and the users of Part IV scores.

FUNDING AND CONFLICTS OF INTEREST

This work was funded internally. The authors are employees of the NBCE.

About the Authors

Igor Himelfarb is the director of psychometrics and research at the National Board of Chiropractic Examiners (901 54th Avenue, Greeley, CO 80634; ihimelfarb@nbce.org). Margaret Seron is a consultant for the National Board of Chiropractic Examiners (901 54th Avenue, Greeley, CO 80634; pegseron@comcast.net). John Hyland is a consultant for the National Board of Chiropractic Examiners (901 54th Avenue, Greeley, CO 80634; jhyland@nbce.org). Andrew Gow is the director of practical testing, research, and development at the National Board of Chiropractic Examiners (901 54th Avenue, Greeley, CO 80634; agow@nbce.org). Nai-En Tang is a data analyst at the National Board of Chiropractic Examiners (901 54th Avenue, Greeley, CO 80634; ntang@nbce.org). Meghan Dukes is a chiropractic specialist at the National Board of Chiropractic Examiners (901 54th Avenue, Greeley, CO 80634; mdukes@nbce.org). Margaret Smith is senior data analyst 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 February 18, 2019; revised May 16 and June 24, 2019; and accepted July 20, 2019.

Author Contributions

Concept development: IH, JH, MS, MD, MF. Design: IH, NT. Supervision: IH, AG, JH. Data collection/processing: MS, MF, NT. Analysis/interpretation: IH. Literature search: JH, MF, NT, IH. Writing: IH, JH, NT, AG. Critical review: JH, AG, MD, MF, MS, NT.

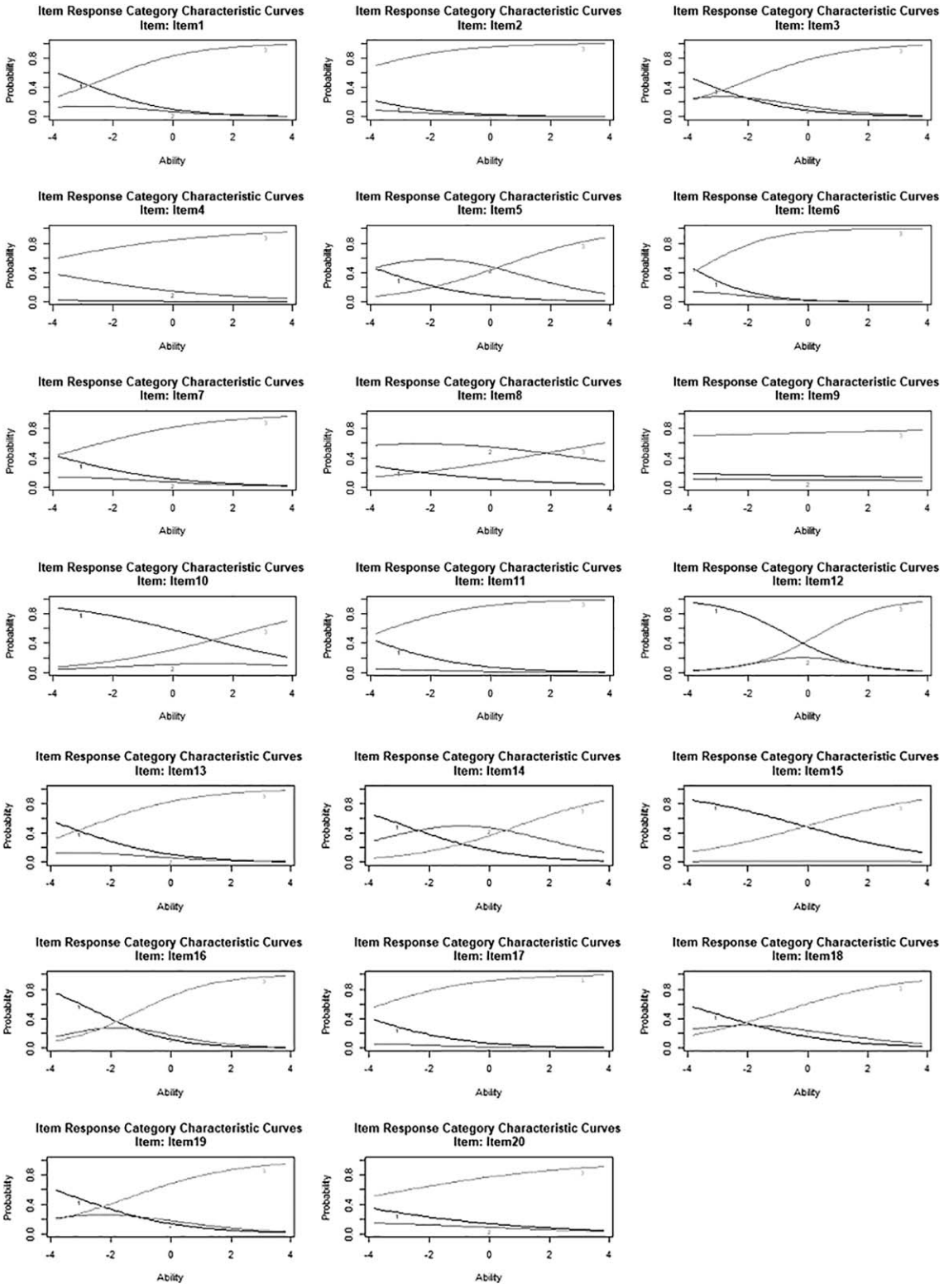

Figure 3 - Item characteristic curves for Form 1 administered in November 2018.

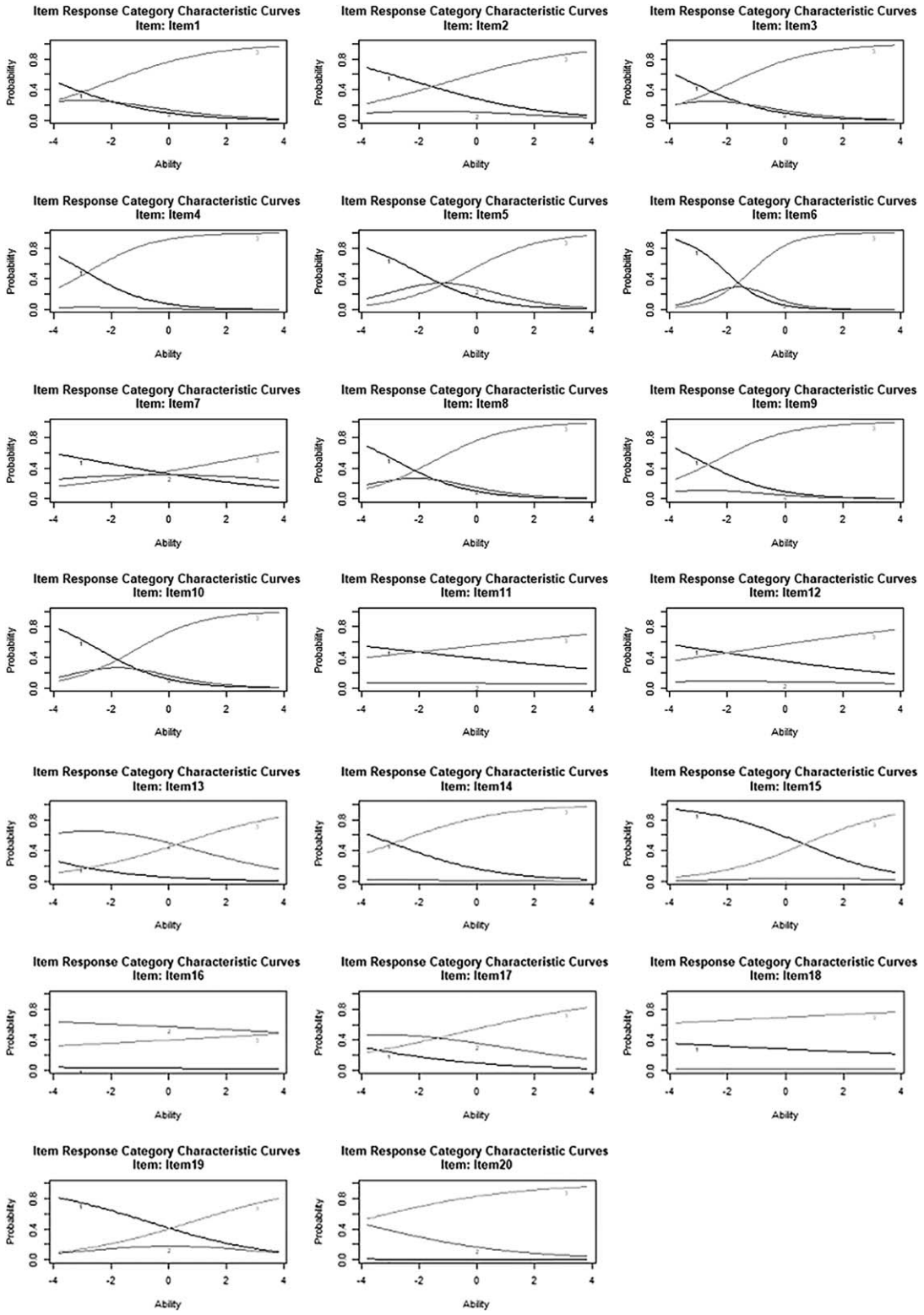

Figure 4 - Item characteristic curves for Form 2 administered in November 2018.

Table 9 - Number of Examinees in Each Administration by Ability Level [a]

Theta Range	November All	November Norming	May All	May Norming
−3 to −2.51	0	0	1	1
−2.5 to −2.01	3	3	5	3
−2 to −1.51	14	12	25	15
−1.5 to −1.01	77	59	97	70
−1 to −0.51	178	147	238	174
−0.5 to 0.01	297	263	346	283
0.0 to 0.49	330	305	346	286
0.5 to .99	261	247	232	201
1.0 to 1.49	122	117	93	80
1.5 to 1.99	16	16	42	39
2.0 to 2.49	0	0	0	0
2.5 to 3.0	0	0	0	0

[a] Norming group consists of nonaccommodated first-time test takers
The ability range is −3.0 to 3.0.

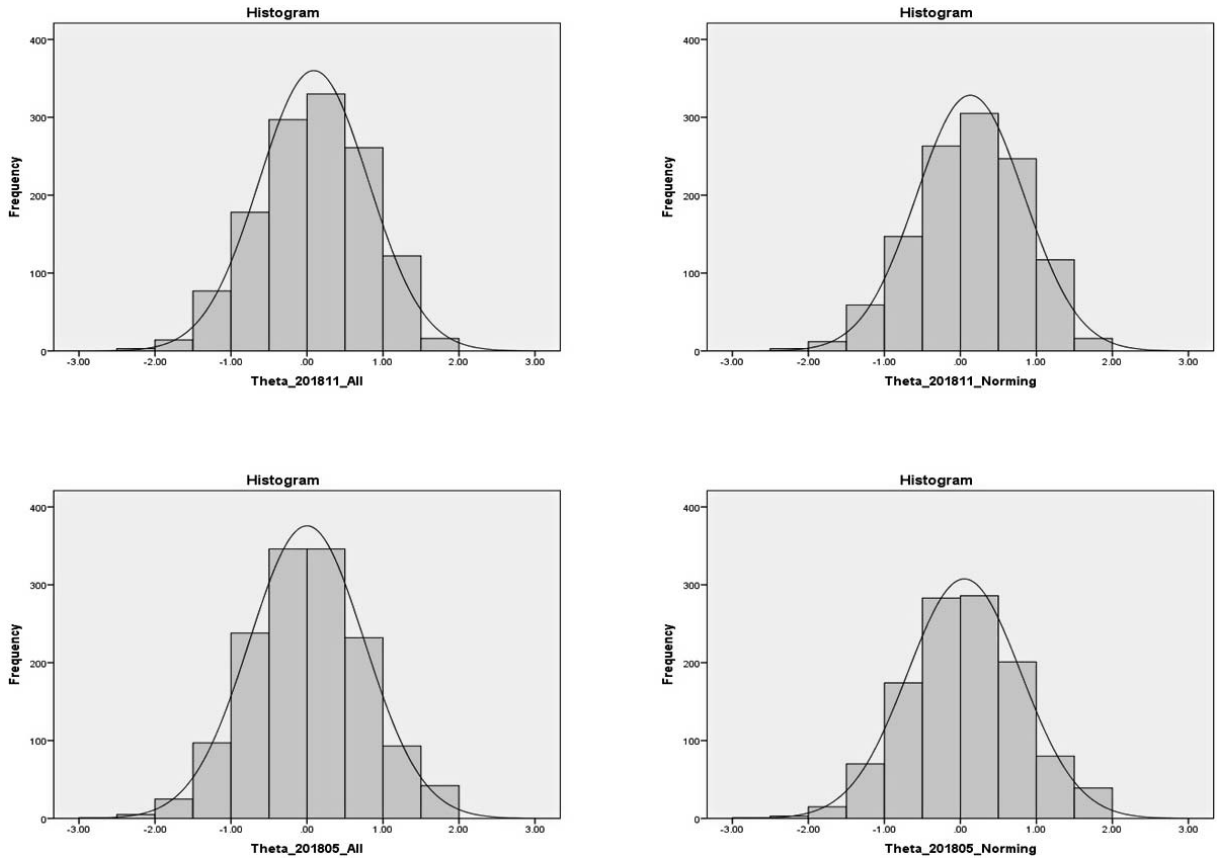

Figure 5 - Frequency distributions of examinees according to ability levels.

REFERENCES

1. Christensen MG, Hyland JK, Goertz CM, Kollasch MW. *Practice Analysis of Chiropractic 2015*. Greeley, CO: National Board of Chiropractic Examiners; 2015.
2. Hyland JK, Seron MA. Use of digital imaging in chiropractic education and practice in the United States. *J Chiropr Educ*. 2016;30(2):e–168.
3. Townsend P, Seron MA, Hyland JK, Fisher M, Kim J. Investigating the use of digital imaging to assess the radiology interpretation skills of graduating chiropractors on the NBCE part IV practical examination. *J Chiropr Educ*. 2017;31(1):78.
4. Merritt CRB, Gerdeman A, Rovinelli R, Capp MP. American Board of Radiology Computer Test Center. *J Digit Imaging*. 2000;13(2):56–58.

5. Mancall EL, Bashook PG, Dockery JL. *Computer-Based Examinations for Board Certification*. Chicago, IL: American Board of Medical Specialties; 1996.
6. Bansal G. Digital radiography. A comparison with modern conventional imaging. *Postgrad Med J*. 2006; 82:425–428.
7. Siegel E, Reiner B, Abiri M. The filmless radiology reading room: a survey of established picture archiving and communication system sites. *J Digit Imaging*. 2000; 13:22–23.
8. Herrmann TL, Fauber TL, Gill J, et al. Best practices in digital radiography. https://www.asrt.org/docs/default-source/research/whitepapers/asrt12_bstpracdigradwhp_final.pdf?sfvrsn= 743d0370_10. Accessed February 7, 2019.
9. Andrich D. A rating formulation for ordered response categories. *Psychometrika*. 1978;43:561–573.
10. Masters GN. A Rasch model for partial credit scoring. *Psychometrika*. 1982;47:149–174.
11. Muraki E. A generalized partial credit model: application of an E-M algorithm. *Appl Psychol Meas*. 1992;16: 159–176.
12. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monogr Suppl*. 1969;17:1–100.
13. Baker JG, Rounds JB, Zevon MA. A comparison of graded response and Rasch partial credit models with subjective well-being. *J Educ Behav Stat*. 2000;25(3): 253–270.
14. de Ayala RJ. *The Theory and Practice of Item Response Theory*. New York: Guilford Press; 2009.
15. Thissen D, Steinberg L. A taxonomy of item response models. *Psychometrika*. 1986;51:567–577.
16. Ostini R, Nering ML. *Polytomous Item Response Theory Models*. Thousand Oaks, CA: Sage; 2005.
17. Cook KF, Dodd BG, Fitzpatrick SJ. A comparison of three polytomous item response theory models in the context of testlet scoring. *J Outcome Meas*. 1999;3(1): 1–20.
18. Naumenko O. *Comparison of Various Polytomous Item Response Theory Modeling Approaches for Task-Based Simulation CPA Exam Data*. Greensboro: University of North Carolina; 2014. https://www.aicpa.org/BecomeACPA/CPAExam/PsychometricsandScoring/TechnicalReports/DownloadableDocuments/Naumenko_polytomous_2014.pdf.
19. Maydeu-Olivares A. Further empirical results on parametric versus non-parametric IRT modeling of Likert-type personality data. *Multivariate Behav Res*. 2005;(40):261–279.
20. 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.
21. 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.
22. Guiliksen H. *Theory of Mental Tests*. New York, NY: John Wiley & Sons; 1950.
23. Crano WD, Brewer MB, Lac A. *Principles and Methods of Social Research*. 3rd ed. New York, NY: Routledge; 2014.
24. Crocker L, Algina J. *Introduction to Classical and Modern Test Theory*. New York, NY: Harcourt Brace Jovanovich College Publishers; 1986.
25. Haertel EH. Reliability. In: *Educational Measurement*. 4th ed. Westport, CT: American Council on Education; 2006:65–111.
26. Kerlinger FN, Lee HB. *Foundations of Behavioral Research*. 4th ed. Orlando, FL: Harcourt Brace Jovanovich College Publishers; 2000.
27. Olson U, Drasgow F, Dorans NJ. The polyserial correlation coefficient. *Psychometrika*. 1982;47(3):337–347.
28. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.r-project.org.
29. Fletcher TD. Package "psychometric." St Louis University, St Louis, MO; 2015.
30. Stevens SS. On the theory of scales measurement. *Science*. 1946;103(2684):677–680.
31. Yen W, Fitzpatrick AR. Item response theory. In: *Educational Measurement*. 4th ed. Westport, CT: Praeger Publishers; 2006:111–153.
32. Hambleton RK, Swaminathan H. *Item Response Theory: Principles and Applications*. Boston, MA: Kluwer; 1996.
33. Agresti A. *An Introduction to Categorical Data Analysis*. 3rd ed. Hoboken, NJ: John Wiley & Sons; 2018.
34. Raykov T, Marcoulides G. *Introduction to Psychometric Theory*. New York, NY: Taylor and Francis Group; 2011.
35. Thissen D, Orlando M. Item response theory for items scored in two categories. In: Thissen D, Wainer H, eds. *Test Scoring*. Hillsdale, NJ: Lawrence Erlbaum Associates; 2001:73–140.
36. Yen WM. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Appl Psychol Meas*. 1984;8:125–145.
37. Yen WM. Scaling performance assessments: strategies for managing local item dependence. *J Educ Meas*. 1993;30:187–213.
38. Nering ML, Ostini R. The general graded response model. In: *Handbook of Polytomous Item Response Theory Models*. New York, NY: Routledge; 2010:77–108.
39. Baker FB. *The Basics of Item Response Theory*. 2nd ed. Washington, DC: ERIC Clearinghouse on Assessment and Evaluation; 2001.
40. Huynh H. Computation and statistical inference for decision consistency indexes based on the Rasch model. *J Educ Behav Stat*. 1990;15(4):353–368.
41. Cai L, Thissen D, du Toit SHC. *IRTPRO for Windows*. Lincolnwood, IL: Scientific Software International; 2011.
42. Rizopoulos D. ltm: an R package for latent variable modeling and item response analysis. *J Stat Softw*. 2006;17(5):1–25.
43. Kang T, Chen TT. Performance of the generalized *S-X*^{2} item fit index for polytomous IRT models. *J Educ Meas*. 2008;45(4):391–406.