0008-3194/96/19–27/$2.00/©JCCA 1996

The reliability and validity of the Canadian Chiropractic Examining Board examinations: a nine year longitudinal study (1987–1995) of their psychometric properties

Claudio Violato, PhD*
Anthony Marini, PhD*
Douglas Lawson, DC**
Murray McEwen, DC**

Data derived from the administration of the Canadian Chiropractic Examining Board (CCEB) examinations for a nine year period, 1987–1995, were used to evaluate the reliability and validity of the subject tests: anatomy, chemistry, chiropractic practice, diagnosis and symptomatology, microbiology and public health, neurology, pathology, physiology, and x-ray interpretation and physics. Nearly two-thousand candidates from more than eighteen chiropractic colleges have written the CCEB exams over this nine year period. The results indicate that the exams have adequate alpha reliabilities (.69 to .80) and theoretically appropriate statistical properties and item characteristics. There is also substantial evidence of content validity. Results from stepwise multiple regression and factor analyses provided evidence for the criterion-related and construct validity of the exams. The implication of these results for the continued refinement and development of the CCEB exams, together with suggestions for on-going research of their reliability and validity, are discussed.
(JCCA 1996; 40(1):19–27)

KEY WORDS: Board examinations, psychometric analyses, reliability and validity, chiropractic.

Données tirées de la commission d'annotation du Canadian Chiropractic Examining Board (CCEB) sur une période de neuf ans (1987 à 1995). Ces données ont servi à évaluer la fiabilité et la validité des examens des sujets : anatomie, chimie, pratique chiropratique, diagnostic et symptomatologie, microbiologie et santé publique, neurologie, pathologie, physiologie, interprétation des radiographies et physique. Environ deux mille personnes de plus de dix-huit collèges de chiropractie ont pris part aux examens du CCEB pendant cette période de neuf ans. Les résultats indiquent que les examens présentent des fiabilités alpha adéquates (de 0,69 à 0,80) et des propriétés statistiques et des caractéristiques d'articles appropriées théoriquement. On trouve également des preuves significatives de la validité du contenu. Les résultats de l'analyse de régression multiple séquentielle et de l'analyse factorielle ont apporté des indications de la validité du test critériel et de la validité conceptuelle des examens. Il s'ensuit une discussion des répercussions de ces résultats afin d'assurer un raffinement et une élaboration soutenus des examens du CCEB, de même que des suggestions pour une recherche continue de leur fiabilité et de leur validité.
(JCCA 1996; 40(1):19–27)

MOTS CLÉS : Commission d'annotation, analyses psychométriques, fiabilité et validité, chiropractie.

* University of Calgary and Edumetrics Ltd.
** Canadian Chiropractic Examining Board.
Correspondence to Dr. Claudio Violato,
Department of Educational Psychology, University of Calgary,
2500 University Drive N.W., Calgary, Alberta T2N 1N4.
Telephone: (403) 220-7296;
Fax: (403) 282-9244; e-mail: violato@acs.ucalgary.ca

Introduction

The Canadian Chiropractic Examining Board (CCEB) conducts what are called "high stakes" exams. These are exams whose results are the basis of very important decisions about the candidates who take them. Examinations in law, medicine, and psychology, for example, which candidates must pass in order to receive a license to practice in

their profession are also high stakes exams (Klass, 1994).[1] Failure on these exams prevents a candidate from achieving a license and therefore practice in their desired profession. College admission tests such as the Law School Admission Test and the Medical College Admission Test are also examples of high stakes exams – performance on them may facilitate or hinder a student from admission to the college.

In Canada, aspiring chiropractors must successfully pass the CCEB exams before they are even eligible to take provincial licensure exams (alsohigh stakes exams). These types of high stakes exams frequently undergo legal challenges (Cavanaugh, 1991; Hambleton, 1995; Pyburn, 1990; Smith and Hambleton, 1990).[2,3,4,5] Litigation over the last two decades makes it clear that the Courts are interested in carefully scrutinizing the procedures used in the development of tests and requiring formal investigations of the validity (the extent to which the tests measure what they are intended to measure) and reliability (the consistency or precision with which the tests measure) of high stakes tests. In addition to the legal requirements of test validation, there isalso the need to develop and validate examinations that are consistent with the Standards for Educational and Psychological Testing, Code of Fair Testing Practices, and the Criteria for Approval of Certification Programs. These latter three documents have been developed and are endorsed by major professional associations in both Canada (e.g., Canadian Psychological Association, Canadian Society for the Study of Education) and the United States (e.g., American Psychological Association, National Council for Measurement in Education, American Educational Research Association) that regulate the professional use of educational and psychological tests.

Besides the legal and professional requirements that high stakes examinations be developed and validated according to the highest standards of testing practices, examining boards have an ethical and moral obligation to undertake such development and validation studies (Hambleton, 1995; Pyburn, 1990; Smith & Hambleton, 1990)[3,4,5]. Because a license confers the legal right to engage in a regulated activity, it is economically very significant. Without the requisite license, a person who attempts to engage in a licensed occupation violates federal and provincial law. It is therefore incumbent upon regulatory testing and licensing boards to provide exams of the highest standards of validity and reliability so as to provide the fairest possible examinations to the candidates, and protect the public from incompetent or unqualified practitioners.

Formal investigations of the reliability and validity of high stakes exams in medicine, for example, are common. In Canada there have been several investigations of the validity and reliability of the Medical Council of Canada (MCC) licensing examinations (e.g., Baumber, 1992; Maguire, Skakun and Harley, 1992; Page, Bordage and Allen, 1995).[6,7,8] Similar studies have been conducted in the United States with the National Board of Medical Examiners (NBME) Subject Examinations and the more recent United States Medical Licensing Examinations (Elam and Johnson, 1994; Mennin, Friedman, Skipper, Kalishman and Snyder, 1993; Williams, 1993).[9,10,11] Similar validation studies have been conducted in nursing (e.g., Bersky and Yocom, 1994)[12] and pharmacy (Fielding, Page, Schulzer, Rogers and O'Byrne, 1992).[13]

The foregoing discussion suggests that a formal study of the reliability and validity of the CCEB exams is warranted. Accordingly, the main purpose of the present study was to investigate the reliability and validity of the CCEB examinations using the broadest data base available: from 1987 to the present (1995). Such data and a review of the procedures used to create the CCEB subject exams will address issues of the reliability, validity and other psychometric properties of the exams.

Method

Subjects

The data from a total of 1,906 candidates (1,395 men – 73%; 511 women – 27%) who wrote the CCEB exams from 1987 to 1995 inclusive, were analyzed in the present study. Candidates from more than eighteen chiropractic colleges wrote the exams over the nine year period of the study (included in these are the Canadian Memorial Chiropractic College, Palmer College of Chiropractic – West, Logan College of Chiropractic, Palmer College of Chiropractic – Davenport as well as others[a]).

The CCEB subject examinations

The current CCEB test battery which consists of nine subject tests (anatomy, chemistry, chiropractic practice, diagnosis and symptomatology, microbiology and public health, neurology, pathology, physiology; x-ray interpreta-

tion and physics) is administered over the course of one week. The main purpose of the CCEB exams is to measure knowledge of the basic sciences of chiropractic at the appropriate cognitive levels. Each test is composed of 75 multiple-choice items that are either selected from an existing computerized item-bank, are created by subject area experts, or by test construction committees. Each test is ultimately created, finalized and edited by test construction committees that are composed of five people – a careful mix of subject area experts, general chiropractors and one Ph.D. level expert in testing. The committees meet during the winter to prepare the exams for administration in April.

Procedures

Data were obtained on all nine of the subject exams of the CCEB test battery for all candidates who wrote these tests from 1987 to 1995 inclusive. Grade averages from the chiropractic colleges from which the candidates graduated were also obtained.

In 1991 the CCEB exams underwent substantial changes – a complete multiple-choice format, implementation of test construction committees, computerized scoring, and criterion-referenced interpretations. Accordingly, the results reported below are of (1) the psychometric and descriptive characteristic conducted on the CCEB exams administered over the four year period, 1992 to 1995, and (2) the analyses addressing the criterion related and construct validity of the exams administered over the nine year period, 1987–1995. The latter were principally multivariate analyses (factor analysis and multiple regression) of test scores and grade averages.

Results

The results are reported under two general headings: (1) Psychometric and Descriptive Characteristics, and (2) Validity Evidence. The second heading is further subdivided into evidence of (1) content validity, (2) criterion-related validity, and (3) construct validity.

Psychometric and descriptive characteristics

Several descriptive characteristics (mean, median, standard deviation) and psychometric properties (alpha reliability, standard error of measurement – SEM, item difficulty and item discrimination) of the nine exams administered over the four year period, 1992 to 1995, are summarized in Table 1. From these results it can be seen that two indicators of central tendency (mean and median) are very stable across the nine exams. The grand mean is the mean for each exam computed over the four year period, as is the grand median. The homogeneous standard deviations for both measures across all of the exams (see Table 1) indicate high stability of the exams across the four year periods as do the standard deviations themselves. The mean skewness of the exams was near zero (0.02), and together with the values of the grand mean and median, indicate a near normal distribution.

The summary of psychometric properties of the exams (Table 1) indicate that they are appropriate. Alpha reliability coefficient is a measure of internal consistency – it summarizes the precision that test items measure in the expected direction (Cronbach, 1951; 1990).[14,15] As can be seen in Table 1, alpha coefficients are in the adequate to very good range (.69 to .80). Their small standard deviations (approximately .05) are indicative of stable reliabilities across years and exams. The standard errors of measurement (SEM), which are another method besides reliability to express error on the tests, also are stable and quite small. These results together indicate that the tests have good internal consistency reliabilities.

Item difficulty is a statistic that summarizes the ease or conversely difficulty of an item. A difficulty index of .80 or greater (.80 or a greater proportion of the candidates got the item correct) suggests the item is easy, while a difficulty index of .30 or less suggests the item is quite difficult. As can be seen from Table 1, the mean difficulty indices (computed across all items for each exam over the four year period) are in the "average" range (.67 to .69). The standard deviations of these indices range from .10 to .12. This means that two-thirds of the item difficulty indices fall within .55 to .79 – optimal difficulties for enhancing reliability (Gronlund and Linn, 1990; Violato, McDougall and Marini, 1992).[16,17] Less than 3% of the items fall within the very difficult range (i.e., difficulty < .30) and 3% in the very easy range (i.e., difficulty > .90). Indeed, 95% of the items fall within the average difficulty (including moderately easy items – .80 to .89 – and moderately difficult items – .39 to .50) range. These are near optimal difficulty ranges for psychometric purposes (Anastasi, 1988; Cronbach, 1990; Violato et al., 1992).[18,14,17]

Item discrimination is another statistic that attests to the CCEB exams' respectable psychometric properties

Table 1
Summary of Descriptive Statistics and Psychometric Properties of the Nine Subject Exams
(n = 1,053; 1992–1995)

Subject Exam	Grand Mean (%)	Grand Median (%)	Mean alpha Reliability	Mean SEM[a]	Mean Item Difficulty	Mean Item Discrimination
1. ANAT	67.7(11.1)[b]	67.2(11.0)	.80(.03)	3.58(.10)	.67(.11)	.36(.13)
2. CHEM	69.0(11.2)	68.7(12.1)	.70(.05)	3.54(.15)	.69(.11)	.31(.12)
3. CHIRO	67.9(11.0)	67.2(11.7)	.68(.09)	3.66(.04)	.68(.10)	.30(.15)
4. MICRO	68.3(11.0)	69.1(12.0)	.77(.03)	3.54(.10)	.68(.11)	.35(.12)
5. XRAY	67.7(11.0)	67.4(11.4)	.73(.04)	3.62(.07)	.67(.11)	.33(.12)
6. DIAG	68.9(10.9)	68.5(11.2)	.71(.08)	3.40(.08)	.69(.10)	.34(.15)
7. PATH	67.8(11.1)	67.9(11.9)	.74(.07)	3.44(.07)	.68(.12)	.33(.12)
8. PHYS	68.3(11.2)	68.1(10.9)	.69(.06)	3.49(.09)	.68(.11)	.30(.12)
9. NEUR	68.3(11.1)	68.5(11.0)	.70(.04)	3.44(.06)	.68(.11)	.32(.13)

[a]SEM = Standard error of measurement
[b]standard deviation

Table 2
Intercorrelation of Nine Subject Tests of the CCEB and GPA
(n = 1,906; 1987–1995)

Test										
1. ANAT	1.00									
2. PHYS	0.70	1.00								
3. CHIR	0.64	0.65	1.00							
4. DIAG	0.65	0.63	0.64	1.00						
5. MICR	0.66	0.67	0.64	0.66	1.00					
6. NEUR	0.72	0.69	0.66	0.65	0.68	1.00				
7. PATH	0.70	0.71	0.67	0.68	0.71	0.74	1.00			
8. XRAY	0.67	0.67	0.66	0.66	0.66	0.70	0.72	1.00		
9. CHEM	0.66	0.69	0.60	0.63	0.67	0.65	0.68	0.65	1.00	
10. GPA	0.44	0.44	0.38	0.45	0.43	0.41	0.42	0.43	0.47	1.00

(Table 1). The discrimination index is a measure of the extent to which an item discriminates between top performers and poor performers. An index of zero indicates that there is no discrimination between top and bottom performers. A positive discrimination indicates the item is discriminating in the desired direction. The mean discriminations (computed across all items over a four year period for each exam) are positive and in the adequate range (Hopkins, Stanley, and Hopkins, 1990).[19] In summary, the descriptive and psychometric properties of the CCEB exams are appropriate.

Validity evidence

Three types of validity evidence are addressed in this section: (1) content, (2) criterion-related, and (3) construct.

Content validity. Content validity concerns the extent to which a test adequately samples the domain of measurement – the content domain and relevant cognitive processes. The content domain is the subject matter in question (e.g., chemistry) while cognitive processes refers to level of knowledge or understanding that is assessed (e.g., rote knowledge, comprehension, application, etc.). Content validity may be most directly enhanced through the use of a table of specifications. The table of specifications is a plan which specifies the content areas to be tested as well as the cognitive outcomes that are to be measured (Violato et al., 1992).[17]

All of the CCEB exams are constructed based on a detailed table of specifications which specifies the percentage of emphasis of each subcontent area of the whole test (e.g., Anatomy/Spine/Lumbar = 14%) and cognitive outcomes (e.g., comprehension, applications). Additionally, since each item is created, revised and edited according to rigorous standards of technical quality (Haladyna, 1994; Haladyna and Downing, 1989; Marini and Violato, 1995),[20,21,22] and is item analyzed after administration, content validity is enhanced. These procedures for creating the tests, therefore, maximize content validity.

Criterion-related validity. When the interest is in how performance on tests correlates with performance on some other criterion, we are concerned about criterion-related validity. In the present analysis, the nine subject tests over the nine year period, 1987 to 1995, were evaluated for criterion-related validity by employing simple intercorrelations (Table 2) and multiple regression analyses (Table 3).

The Pearson correlations in Table 2 clearly show patterns of strong correlations between all of the subject tests (i.e., $r > .60$) and moderate correlations of the tests with GPA ($r > .38$). (All grades were converted to a four-point scale – GPA – which is in common use.) These correlations are evidence of criterion-related validity as the tests all result in strong intercorrelations. The pattern of correlations with GPA provide further evidence of criterion-related validity as test performance on CCEB exams correspond to GPA assigned at chiropractic colleges.

To evaluate further the criterion-related validity of the CCEB, a stepwise multiple regression was conducted with GPA as the dependent variable and the nine subject tests as independent variables. The results from this analysis are reported in Table 3.

Multiple regression is a statistical procedure for analysing the multivariate interrelationship between a number of independent variables with each other, and these variables with a dependent or criterion variable. Accordingly, this permits a more complete and full understanding of the complex interrelations between a set of variables than do single univariate correlations. The stepwise procedure summarized in Table 3 allows for entry of each variable into the analysis one at-a-time so as to evaluate its contribution to the multiple regression. The order of entry of the independent variables is based on their statistical significance – the most significant variable is first, the next second, and so on.

As can been seen from the results in Table 3, Chemistry which has the largest simple correlation ($r = .47$; $p < .01$) with GPA (also Table 2), is first. Since there is only one variable in the equation at this point (step 1), the simple correlation is the same as the multiple r (correlation between all independent and the dependent variables). The r^2 is the proportion of variance accounted for in the dependent variable (GPA) by all of the independent variables. The change in r^2 indicates the increase in variance accounted for by the inclusion of that variable. Beta weights are used to weigh each variable in the prediction equation.

Step 2 in the regression is the inclusion of Diagnosis and Symptomatology (Multiple $r = .51$; $p < .01$). The increase in variance accounted for is modest (r^2 change = .04). The next three steps include Anatomy ($p < .05$), Physiology ($p < .05$), and Microbiology and Public Health ($p < .05$). The remaining tests (Chiropractic Practice, X-ray, Neurology, Pathology) failed to make any further contributions to

Table 3
Stepwise Multiple Regression of Nine Subject Exams and GPA
(n = 1,906; 1987–1995)

Step	Variable	Multiple r	r^2	r^2 change	Beta Weights
1	CHEM**	.47	.22	.22	.20
2	DIAG**	.51	.26	.04	.15
3	ANAT*	.52	.27	.01	.10
4	PHYS*	.53	.28	.01	.08
5	MICR*	.54	.29	.01	.08
6	CHIR	.54	.30	.01	.01
7	XRAY	.54	.30	.00	.06
8	NEUR	.54	.30	.00	.12
9	PATH	.54	.30	.00	−.84
					Constant = 31.86

** p<.01; * p<.05

Table 4
Factor Structure Matrix Obliquely Rotated to the Kaiser Normalized Oblimin Criterion
(n = 1,906; 1987–1995)

Variable	Factor 1 Chiropractic Science	Factor 2 Scholastic Aptitude
1. ANAT	.84	.43
2. PHYS	.85	.43
3. CHIR	.82	.33
4. DIAG	.82	.46
5. MICR	.84	.43
6. NEUR	.86	.38
7. PATH	.88	.39
8. XRAY	.85	.40
9. CHEM	.82	.52
10. GPA	.50	.99
Percent of Variance	66.5	7.3
	Total Variance = 73.8	
Correlation between Factor 1 and Factor 2 = 0.49		

the equation or reach statistical significance. This is evident from the data in Table 3.

These results from the stepwise multiple regression further help to clarify the interrelationships among the subject tests and between GPA as the criterion. The subject tests have statistically significant correlations among themselves and multiple r with GPA. Nearly all of the variation (26%) in GPA, however, is accounted for by the first two steps (Chemistry and Diagnosis and Symptomatology). The remaining tests either make little contribution or none that is significant. This is because these tests are so highly intercorrelated (see also Table 2) that they are redundant in the regression equation.

Construct validity. Construct validity focuses on the truth or correctness of a construct (a hypothetical entity or process which is itself not observed) and the instruments that measure it. This type of validity requires the gradual accumulation of information from a variety of sources.

Factor analysis is a multivariate statistical technique that allows for the identification of underlying constructs or latent variables in a data set such as a correlation matrix. As such, it is one method of addressing issues of construct validity of tests. Accordingly, a factor analysis was conducted on the correlation matrix in Table 2.

As a first step, the matrix (Table 2) was decomposed into principal components. Upon inspection of the resulting eigen values and variance accounted for, it was decided to retain two principal components. A close inspection of the correlations in Table 2, indicates that all variables are significantly intercorrelated as we have seen. Accordingly, the two principal components were rotated obliquely (rather than orthogonally) so as to permit their correlation as indicated in the pattern of correlations summarized in Table 2. The factor structure matrix which resulted from the obliquely rotated principal components to the Kaiser normalized Oblimin criterion is summarized in Table 4.

Two factors, Chiropractic Science (Factor 1) and Scholastic Aptitude (Factor 2), emerged from the foregoing analysis. All nine subject tests have very high factor loadings (> .80 see Table 4) on Factor 1. This factor was named Chiropractic Science because it clearly underlies performance on the CCEB subject tests which are intended to measure the content and processes of the science knowledge of chiropractic. Factor 2 was named Scholastic Aptitude because it has a very high loading from GPA (.99 see Table 4) and moderate loadings from several subject tests (.43 to .52). The very high loading from GPA with small or moderate loadings from the subject tests clearly indicate that the second factor is separate and theoretically meaningful. As might be expected theoretically, there is a significant correlation between Chiropractic Science and Scholastic Aptitude (r = .49). The total percentage of variance in the data accounted for by this two factor solution is 73.8 (see Table 4). The results of the factor analysis clearly provide evidence for the construct validity of the CCEB subject exams.

Discussion

The main results of the present study may be summarized as follows: (1) the descriptive and psychometric results from the exams over the four year period, 1992 to 1995, indicate that the CCEB exams have adequate reliability, item characteristics and distributional properties, (2) considerable care is taken to enhance the content validity of the exams, and (3) there is evidence of both criterion-related and construct validity of the CCEB exams as indicated by the analyses of the exams and GPA over the nine year period, 1987 to 1995.

The average reliabilities of the exams are generally in the adequate range. With continued refinement of the exams by applying vigorous quality control the reliabilities should continue to improve eventually reaching values in the .90s. Similarly, while the item discriminations are also in the adequate to good range, further work should increase the discriminations. In particular, enlarging the item bank to include more items at higher levels of cognitive complexity (comprehension, application, synthesis) should directly increase average item discrimination (Gronlund and Linn, 1990).[16] So while the psychometric properties of the exams are currently appropriate, continued refinement and work should result in further improvements in these properties.

There is considerable attention given to the content validity of the CCEB exams. This effort is focused primarily on two matters: (1) the development and use of detailed tables of specifications for item selection, and (2) the application of high standards of item creation, editing and refinement. Detailed tables of specifications is probably the most effective method of enhancing content validity (Hopkins et al., 1990; Violato et al., 1992).[19,17] The tables of specifications themselves, however, need to be continually validated with ongoing research such as the scope of practice of the profession, practice analysis and role deline-

ation of chiropractors (Kane, Miller, Trine, Becker and Carson, 1995; Lunz, Stahl, and James, 1989; Smith and Hambelton, 1990; Standards for Educational and Psychological Testing, 1995),[23,24,5,25] Lunz et al. (1989),[24] for example, have developed a quantitative method for linking role delineation to test specifications rather than relying on expert intuition only. Using survey procedures, Delphi techniques and focus groups, tables of specifications of both content and cognitive complexity can be derived by statistical methods. Currently, the CCEB is funding a three year study whose main purpose is to develop tables of specifications for the CCEB exams employing these procedures.

The second major concern for content validity is to create items of the highest technical quality. The CCEB employs experts in testing to train experts in content areas to create items of the highest calibre. Using this approach, a team of item writers from several major universities across Canada (UBC, Calgary, Toronto) has been developed. These people are either graduate students (MSc or PhD) or completed experts (PhD or MD) in relevant sciences (anatomy, pathology, microbiology, neurology, physiology, chemistry, pharmacology, nutritional sciences, exercise physiology). Chiropractors write items for the chiropractic practice and x-ray interpretation and physics exams, and also evaluate the clinical relevance of the basic sciences items. Each content specialist attends an item construction workshop where they are taught the techniques of sound test item construction (Haladyna, 1994; Haladyna and Downing, 1989; Marini and Violato, 1995),[20,21,22] The item writers are provided with feedback on their initial efforts and their work is subject to quality control. Ultimately, these items writers produce well constructed and content accurate test questions. Using this approach, nearly 2,000 new items have been added to the CCEB item bank during the past year. These items will be administered eventually, item-analyzed and revised based on the results. These substantial efforts both in developing the tables of specifications and items themselves are state of the art methods that enhance the content validity of the CCEB exams.

The results from the correlational analysis and regression analyses both provide evidence of criterion-related validity while the factor analysis provides evidence of construct validity. In the multiple regression, the patterns of multiple correlations and the patterns of simple correlations from the correlational analysis follow theoretical expectations: correlations among the subject tests are high and larger than the correlations with GPA. Nevertheless, the criterion-related validity of the CCEB exams should continue to be investigated and researched using other relevant criteria. These might include, for example, such criteria as performance on licensing clinical competency exams, performance on patient management problems, efficacy in clinical practice, and so on (Association of American Medical Colleges, 1993).[26]

While the present factor analysis does provide some evidence of construct validity, further work is required in this area. The process of construct validation is a complex, multifaceted one that requires multiple research methods, types of data and various investigative approaches (Anastasi, 1988; Cronbach, 1990).[18,15] No single study is sufficient to establish construct validity – rather an ongoing programatic research enterprise is required. The aforementioned three year CCEB study will provide more precise data of the scope of practice of chiropractors than now exists, as well as a clearer understanding of role delineation. Such data will prove useful to study further the construct validity of the CCEB exams. Meanwhile, there is a cohesive factor – Chiropractic Science – which is separate but related to (i.e., correlated with) the second basic construct – Scholastic Aptitude. In short, the CCEB exams are not merely another measure of scholastic aptitude – they measure the basic sciences that are fundamental to chiropractic.

Summary and conclusions

The results of the present study show that the CCEB subject exams have appropriate psychometric properties. Moreover, they indicated that there is substantial evidence of content validity, and some evidence of criterion- related and construct validity. Further work which has already been initiated is likely to enhance both the content and construct validity of the exams. There are also currently research plans to investigate further the criterion-related validity employing clinical competency measures as the external criterion. With continued refinement and development, the CCEB exams which already meet high standards, can achieve the highest calibre of reliability and validity that the science of educational and psychological measurement has to offer. In this way, the CCEB is fulfilling not only its legal and technical obligations, but its moral and ethical ones as well.

References

1 Klass DJ. "High-Stakes" testing of medical students using standardized patients. Teaching and Learning in Medicine 1994; 6:28–32.
2 Cavanaugh SH. Response to a legal challenge: Five steps to defensible credentialing examinations. Evaluation and the Health Professions 1991; 14:13–40.
3 Hambleton RK. Setting standards on performance assessments: Technical issues and promising methods. Paper presented at the 103rd Annual Convention of the American Psychological Association, New York, 1995.
4 Pyburn KM. Legal challenge to licensing examinations. Educational Measurement: Issues and Practice 1990; (Winter):5–14.
5 Smith L, Hambleton RK. Content validity studies of licensing examinations. Educational Measurement: Issues and Practice 1990; (Winter):7–10.
6 Baumber J (ed). Objectives of the Medical Council of Canada Qualifying Examination. Ottawa: Medical Council of Canada, 1992.
7 Maguire T, Skakun E, Harley C. Setting standards for multiple-choice items in clinical reasoning. Evaluation and the Health Professions 1992; 15:434–452.
8 Page G, Bordage G, Allen T. Developing key-feature problems and examinations to assess clinical decision-making skills. Academic Medicine 1995; 70:194–201.
9 Elam CL, Johnson MS. NBME Part I versus USMLE Step 1: Predicting scores based on preadmission and medical school performance. Academic Medicine 1994; 69:155.
10 Mennin SP, Friedman M, Skipper B, Kalishman S, Snyder J. Performances on the NBME I, II, and III by medical students in the Problem-based Learning and Conventional tracks at the University of New Mexico. Academic Medicine 1993; 68:616–624.
11 Williams RG. Use of the NBME and USMLE examinations to evaluate medical programs. Academic Medicine 1993; 68:748–752.
12 Bersky AK, Yocom CJ. Computerized clinical simulation testing: Its use for competence assessment in nursing. Nursing and Health Care 1994; 120–127.
13 Fielding DW, Page GG, Schulzer M, Rogers T, O'Byrne CC. Assuring continuing competence: Identification and validation of a practice-based assessment blueprint. American Journal of Pharmaceutical Education 1992; 56:21–19.
14 Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951; 16:297–334.
15 Cronbach LJ. Essentials of psychological testing (5th ed). New York: Macmillan, 1990.
16 Gronlund NE, Linn JR. Measurement and evaluation in teaching (6th ed). New York: Macmillan, 1990.
17 Violato C, McDougall D, Marini A. Educational measurement and evaluation. Dubuque, IA: Kendall/Hunt, 1992.
18 Anastasi A. Psychological testing (6th ed). New York: Macmillan, 1988.
19 Hopkins KD, Stanley JC, Hopkins BR. Educational and psychological measurement and evaluation (7th ed.). Englewood Cliffs, NJ: Prentice Hall.
20 Haladyna TM. Developing and validating multiple-choice test items. Hillsdale, NJ: Erlbaum, 1994.
21 Haladyna TM, Downing SM. A taxonomy of multiple-choice test items. Applied Measurement in Education 1989; 2:37–50.
22 Marini A, Violato C. A technical guide to writing multiple-choice items. Calgary: Laboratory for Educational Measurement and Evaluation, 1995.
23 Kane MT, Miller T, Trine M, Becker C, Carson K. The precision of practice analysis results in the professions. Evaluation and the Health Professions 1995; 18:29–50.
24 Lunz ME, Stahl JA, James K. Content validity revisited. Evaluation and the Health Professors. 1989; 12:192–206.
25 Standards for Educational and Psychological Testing. Washington DC: American Psychological Association, 1995.
26 Association of American Medical Colleges. Trends plus: U.S. medical school applicants, matriculants, graduates, 1992. Washington DC: AAMC, 1993.

Endnotes

[a] Canadian Memorial Chiropractic College; Palmer College of Chiropractic – Davenport; Palmer College of Chiropractic – West; Anglo-European College of Chiropractors; Northwestern College of Chiropractic; Logan College of Chiropractic; Los Angeles College of Chiropractic; Western States Chiropractic College; Life College of Chiropractic; Life College of Chiropractic – West; Parker College of Chiropractic; Cleveland Chiropractic College; National College of Chiropractic; New York Chiropractic; Northwestern College of Chiropractic; Texas College of Chiropractic; Philip Institute of Technology, School of Chiropractic and Osteopathy.