ORIGINAL ARTICLE

Examining the validity of chiropractic grade point averages for predicting National Board of Chiropractic Examiners Part I exam scores

Igor Himelfarb, PhD, Bruce L. Shotts, DC, and Andrew R. Gow, DC

ABSTRACT

Objective: The main objective of this study was to evaluate the validity of grade point average (GPA) for predicting the National Board of Chiropractic Examiners (NBCE) Part I exam scores using chiropractic GPA.

Methods: Data were collected during the January 2019 computer-based testing administration of the NBCE's Part I exam. The sample size was $n = 2278$ of test takers from 18 domestic and 4 international chiropractic educational institutions. Six regression models were developed and tested to predict the Part I domain scores from chiropractic GPA while controlling for self-reported demographic variables. Residuals from the models were disaggregated by pre–chiropractic GPA.

Results: Chiropractic GPA revealed a positive, statistically significant correlation with sex. The chiropractic GPA was found to be a significant predictor of the Part I domain scores. A different perspective was obtained when residuals (observed minus predicted) were collected and split by the pre–chiropractic GPA. *Very good students* tended to be underpredicted, while *other students* were overpredicted.

Conclusion: This study builds on the cascading evidence from educational literature by providing additional results suggesting that undergraduate (prechiropractic) GPA as well as the GPA obtained in doctor of chiropractic programs are related to the future performance on the NBCE Part I exam. The results provide a first glance at the connection between the standardized test scores, which are often used for instructors' and institutional evaluation and the GPA obtained in a doctor of chiropractic program.

Key Indexing Terms: Chiropractic; Educational Measurement; Regression Analysis; Grade Point Average

INTRODUCTION

In today's culture of educational accountability, and with pressure and support from the US Department of Education and accreditation agencies, educational institutions built instructors' accountability systems to include, among other measures, one or more indicators based on students' performance on standardized tests.[1] The results of tests that were designed and validated to probe the knowledge of examinees in specific subject areas are now used to evaluate instructors' performance, which raises the validity questions. The issue of the validity of test scores used is not new. Forty years ago, Samuel Messick, an American psychologist who devoted his life to validity research, stated that "responsible use of test scores requires that the test user be able to justify the inferences drawn having a cogent rationale for using test scores for the

purpose at hand and for selecting this test over other available assessment procedures."[2,3]

Another great American educational psychologist, Lee Cronbach said that "Validity was once a priestly mystery, a ritual performed behind the scenes, with professional elite as witness and judge. Today it is a public spectacle combining the attractions of chess and mud wrestling."[4]

The *Standards for Educational and Psychological Testing*[5] formally requires 25 validity standards, but some samples are the following:

> A rationale should be presented for each intended interpretation of test scores for a given use, together with a summary of the evidence and theory bearing on the intended interpretation.—Standard 1.2

> If validity for some common or likely interpretation for a given use has not been evaluated, or if such an interpretation is inconsistent with available evidence,

that fact should be made clear and potential users should be strongly cautioned about making unsupported interpretations.—Standard 1.3

If a test score is interpreted for a given use in a way that has not been validated, it is incumbent on the user to justify the new interpretation for that use, providing a rationale and collecting new evidence, if necessary.—Standard 1.4

When interpretation of subscores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretation should be provided. When composite scores are developed, the basis and rationale for arriving at the composites should be given.—Standard 1.14

The NBCE develops, administers, scores, and reports test scores for examinees, state boards, and chiropractic colleges. The scores were initially intended for making inferences about the examinees' competence required to practice chiropractic. However, since NBCE scores are now used to make inferences about instructors teaching courses and institutions, it is important to establish a formal connection between the scores and the inferences made. It will take time and many studies to build this validity argument; however, when it is in place, chiropractic education in the United States will enjoy a warranted system of assessment and accountability.

The NBCE exams are administered to establish minimum competency to practice chiropractic, which is achieved by scaling the cut score of 375 to the minimum competency point on the ability continuum. However, the scores for the examinees are produced on a continuous scale ranging from 125 to 800 ($M = 500$, SD $\approx$ 100) and cover 99% of the ability spectrum.

This study evaluates the validity of the grade point average (GPA) obtained in chiropractic programs (chiropractic GPA) for predicting performance on the domains of the NBCE's Part I examination. The Part I exam consists of 6 domains: general anatomy (GEA), spinal anatomy (SPA), physiology (PHY), chemistry (CHE), pathology (PAT), and microbiology (MIC). The relationship between GPAs and Part I scores may be evaluated through predictive validity studies. According to *Standards*, predictive validity is defined as the degree to which data are predictive of scores obtained at a later time.[5] The predictive value of GPA is typically assessed by performing a regression analysis in which GPA alone or with other variables serve as predictors of criterion scores.[6] Predictive validity studies usually have 2 purposes— they are used to provide evidence of validity and become a part of validity literature and/or they provide results that may be used at the institutional level to determine how heavily GPA should be considered when determining outcomes such as standardized test scores.[6,7]

Literature Review

A literature of GPA as a predictor of academic success is provided here for context. The relationships between GPA, standardized tests, and academic success drew the attention of researchers, educators, and policy makers as early as the mid-20th century. In 1939, Sarbin[8] conducted a regression analysis to evaluate predictors of college grades.[9] Later, several studies examined performance-related outcomes being predicted by admission criteria, including GPA.[10,11] These studies concluded that GPA might be informative for admission or certification decisions. A different study investigated and confirmed predictive validity of GPA for scores obtained in a performance test.[12] Mehrens and Phillips[13] researched the relations between GPA and test scores for teacher licensing decisions. After examining the models, which included students' GPA and standardized test scores as predictors of future performance, researchers found GPA to be informative for licensing decisions. They suggested that if grades are determined accurately, reliably, and fairly, "they should communicate the degree to which individuals have learned course objectives."[13]

Often predictive validity of GPA is studied in conjunction with standardized test scores. Camara and Echternacht[14] examined the utility of SAT scores and GPA in predicting college success. While the study mainly focused on establishing predictive validity for the SAT, the authors revealed that students' GPA is a slightly better predictor of academic success in college. They continued by saying that the best prediction of college success is reached by combining the SAT scores with GPA; however, GPA alone carries about 46% of the predictive weight. The authors advocated for consideration of cumulative GPA rather than focusing on course-specific GPA (for example, GPA obtained in science classes) when predicting future college success. "The rationale for considering cumulative GPA as an indicator of success in college is that it encompasses the entire scholastic performance of a student in college. Cumulative GPA appears to provide a more comprehensive view of student academic performance."[14] (p. 5)

Hu[15] demonstrated that GPA is the best predictor of future academic success. In his study, the author suggested a revision of the college admission index (a linear combination of the high school GPA and the total SAT score) by measuring the correlation between high school GPA and SAT scores with the GPA obtained during the second term in college. The study showed that GPA, as a predictor of future college performance, was reliable and consistent throughout the course of college study. He concluded that GPA is one of the major predispositions related to student academic ability.[15]

Noble and Sawyer[16] used GPA and the ACT composite score to predict different levels of academic success in college. This study included information from 216 institutions across the United States. Logistic regressions were constructed to predict categories of success in college using the GPA and ACT scores. The results showed that a 4.0 high school GPA corresponded to a 3.75 college GPA in most of the institutions. Further, the research showed that students with GPA values of less than 3.0 had little chance of significantly increasing their GPA in college and usually obtained grades lower than B.[16]

Zwick and Himelfarb[6] analyzed 70,812 students from 34 colleges (data were provided by the College Board) to understand the relationships between high school GPA, SAT scores, and college success (measured by 1st-year

college GPA). Three types of regression models were considered: models with just high school GPA as the predictor of college performance; models that combined GPA with SAT scores; and models that included institution quality index in addition to the GPA and SAT scores. High school GPA was found to be the strongest predictor of college success accounting for more than 20% of total variability in the 1st-year college GPA.[6] Zwick[9] speculated why high school GPA consistently emerged as the strongest predictor of 1st-year college GPA. Referring back to Zwick and Himelfarb,[6] she concluded that a "detailed analysis of the interrelationships of grades, test scores, and socioeconomic status (SES) yields a more complex picture (p. 14)" as both GPA and admission test scores are correlated with SES. However, GPA is "less contaminated with SES than are admission test scores."[9]

The studies previously cited all used GPA from previous educational programs to predict GPA in the program for which a student was applying or had been admitted. Standardized test scores encompassed in these analyses were always included as an additional predictor of the GPA. One of the reasons for this inclusion is because these studies were conducted to refine program admission criteria and the standardized test scores were usually obtained from preadmission testing such as the SAT or ACT. The purpose of this study is a bit different. We are trying to establish predictive validity of the GPA in predicting pre–licensure exam scores. Thus, the scores obtained on the NBCE exams serve as dependent variable rather than predictors in our analytic paradigm.

Current Study

Several recent studies have documented the factors related to success rates on board-certification exams. Lower admission requirements, specifically GPA, were found to be associated with lower passing rates on national credential examinations in allied health programs.[17] Furthermore, GPA was identified as the predictive factor of success on the National Council Licensure Examination for registered nurses.[18] Preclinical course grades were found to be predictive of performance on the National Board of Medical Examiners clinical subject exams.[19] Finally, undergraduate GPA and scores from the Psychological Services Bureau Health Occupations Aptitude Examination were found to be predictive of successful performance on the National Board of Dental Hygiene Examination.[20] However, very little is known of the connection between the chiropractic GPA and performance on the NBCE exams, as previous efforts to investigate these relationships suffered from lack of data availability or were limited by available sample sizes.

This study investigates the predictive validity of chiropractic GPA for the NBCE Part I exam, while controlling for the demographic characteristics of the examinees. The following research questions were addressed:

RQ 1: After controlling for demographic characteristics of the sample, is chiropractic GPA predictive of the Part I domain-level exam scores?

RQ 2: How much of the variability in the Part I scores is explained by the chiropractic GPA?

RQ 3: Is the quality of prediction related to the pre–chiropractic GPA?

METHODS

Overview

We used data collected during the 2019 computer-based testing (CBT) administrations of the recently reduced Part I exam. Starting in 2019, new examinees were asked to provide ancillary information during their NBCE application process. This additional information included self-reported demographic characteristics, educational variables, and indicators of each candidate's socioeconomic status. The educational variables encompassed self-reported pre–chiropractic GPA and GPA from a chiropractic program. Our logic was to investigate the predictive validity of the chiropractic GPA while controlling for the variability accounted by the demographic characteristics and other educational variables. The NBCE institutional review board committee granted the study an exemption status from the full review on November 12, 2019.

Data, Sample, and Measures

There were 3 different categories of the Part I test takers in 2019—the first-time examinees who took the new version of the exam delivered on computers; repeaters who took the new version of the exam having to repeat all 6 domains of the test on computer; and single-subject repeat takers who took an old version of the exam (single domain). Only CBT examinees were included in the study. The item responses were extracted from the operational files, while the ancillary information came from the NBCE registration database. The item responses were matched with the ancillary variables according to the unique NBCE examinee identification number. The preliminary files did not contain names or any other examinee personal information. After merging the item responses with the ancillary information, the NBCE identification numbers were deleted, making the data analysis file completely anonymous.

The initial sample size of test takers for whom we had scores was $n = 2278$. However, not all examinees provided valid responses to ancillary items. Therefore, listwise deletion, a method that excludes an entire record from analysis if any single value is missing, was implemented. Following data cleaning procedures, the sample sizes included in the analyses were $n = 1841$ (Model 1), $n = 1867$ (Model 2), $n = 1857$ (Model 3), $n = 1842$ (Model 4), $n = 1839$ (Model 5), and $n = 1852$ (Model 6).

Participants ranged in age from 22 to 61 years old, $M = 27.09$, $SD = 4.41$; 41% were male and 59% female. In terms of ethnicity, 68.2% were White; 8.8% of the sample were Asian/Pacific Islander test takers; 4.5% were African American; 11.7% were Latino; less than 1% were Native American, and 6.1% were of other or mixed ethnicities. The *Other* category was used as the reference group in all statistical analyses involving ethnicity.

Table 1 - Descriptive Statistics for Key Variables Included in the Study

Variable	Mean	SD	N	%
General anatomy	498.72	83.80	1974	—
Spinal anatomy	497.11	104.62	1974	—
Physiology	498.38	105.49	1965	—
Chemistry	498.77	86.21	1948	—
Pathology	498.39	82.68	1945	—
Microbiology	497.62	90.81	1958	—
Age	27.09	4.41	2278	—
Sex				
Male	—	—	938	41%
Female	—	—	1340	59%
Ethnicity				
Asian/Pacific Islander	—	—	201	8.8%
African American	—	—	102	4.5%
Latino	—	—	266	11.7%
Native American	—	—	16	0.7%
White	—	—	1553	68.2%
Other	—	—	140	6.1%
Education				
Some college	—	—	100	4.4%
Associate's degree	—	—	101	4.4%
Bachelor's degree	—	—	1928	84.6%
Master's degree	—	—	91	4.0%
Doctoral degree	—	—	44	1.9%
Unspecified	—	—	14	0.6%
Pre–chiropractic GPA	3.19	0.45	—	—
Chiropractic GPA	3.32	0.35	—	—

Mean and SD are given for continuous variables, while percentages are given for categorical variables.

Pre–chiropractic education was assessed by asking the respondents to provide their highest level of education prior to being admitted to a chiropractic institution. The response options included *some college*, *associate degree*, *bachelor's degree*, *master's degree*, *doctoral degree*, and *unspecified*. Most respondents reported completion of a bachelor's degree (84.6%), followed by master's degree (4.0%), and doctoral degree (1.9%). These statistics are in line with information reported in the 2015 Practice Analysis of Chiropractic[21]: in 2014, the proportion of practicing chiropractors who had obtained a bachelor's or higher degree was 78.8%.

Two questions solicited responses to assess the GPA obtained prior to and during each candidate's chiropractic program. The first question stated, "What was your undergraduate grade point average (GPA)?" referring to pre–chiropractic GPA. The second question asked, "Please state your current grade point average (GPA) from the chiropractic college you are attending." The response option for each question was limited to a number on a 0.0 to 4.0 scale. As presented in Table 1, the average pre–chiropractic GPA was $M = 3.19$, SD = .45; while the average of chiropractic GPA was $M = 3.32$, SD = .35.

Regression Models

When studying a relationship between 2 variables, it is reasonable to assume that these variables are linearly related. With some degree of tolerable error, a line is the best graphical representation of the relationship between 2 normally distributed variables. We tested the assumption of a linear relationship between the GPA and Part I domain scores numerically and graphically and concluded that there was not enough evidence to support the violation of linearity.

To test the linearity assumption, we plotted the expected value for our dependent variables against each independent variable, while holding the other independent variables fixed. All plots produced a straight line. To test for homoscedasticity, we plotted residuals from each regression model against the predictors. No indication of dependency was revealed.

Another assumption we made concerns the normal distribution of the dependent variable; that is, Domain $Score_{ij} \sim N(\mu; \sigma)$, where Domain $Score_{ij}$ is the test score of an examinee i on domain j, $i = 1, 2, \ldots, N$, $j =$ GEA, SPA, PHY, CHE, PAT, and MIC; μ is the mean parameter, and σ^2 is the variance. The assumption of normal distribution was supported by the diagnostic analysis.

To test the outcomes for normality, we examined the values of skewness and kurtosis. We followed a general guidance by Hair et al,[22] that if skewness and kurtosis are less than $|1|$, there is no indication of significant violation of assumed normality. The following statistics were estimated: skewness $= -.03$, kurtosis $= -.20$ for GEA; skewness $= .15$, kurtosis $= -.06$ for SPA; skewness $= -.03$, kurtosis $= -.33$ for PHY; skewness $= .05$, kurtosis $= -.35$ for CHE; skewness $= .13$, kurtosis $= -.02$ for PT; and skewness $= .03$, kurtosis $= -.15$ for MIC. All SE values were .05 for skewness and .11 for kurtosis. Additionally, we examined Q–Q plots, which compare expected and observed values under the assumption of normal distribution.

Next, 6 regression models were estimated and tested:

$$\text{Domain Score}_{ij} = \beta_0 + \beta_1 d(\text{Male})_{ij}$$
$$+ \beta_2 d(\text{Asian/Pacific Islander})_{ij}$$
$$+ \beta_3 d(\text{African American})_{ij} + \beta_4 d(\text{Latino})_{ij}$$
$$+ \beta_5 d(\text{Native American})_{ij} + \beta_6 d(\text{White})_{ij}$$
$$+ \beta_7 (\text{Education})_{ij}$$
$$+ \beta_8 (\text{ChiropracticGPA})_{ij} + e_{ij},$$

where Domain $Score_{ij}$ is the test score of an examinee i on domain j, $i = 1, 2, \ldots, N$, $j =$ GEA, SPA, PHY, CHE, PAT, and MIC; β_0 is the intercept, and $\beta_1, \ldots, \beta_8$ are regression coefficients (slopes) associated with the predictors in the model. The term $d(\cdot)$ is the indicator for the parenthesized predictor; that is, $d = 1$ if the examinee has the characteristic in question and $d = 0$, otherwise. The e_{ij} is the error term representing the difference between the predicted and observed scores. We make the following assumption for the errors:

$$e_{ij} \overset{iid}{\sim} N(0, \sigma^2).$$

The Bonferroni adjustment was implemented to control for a possible Type I error inflation owing to multiple analysis on the same dependent variable.[23] The correction was made to the p values used for evaluation of statistical

significance. We used the following formula to perform Bonferroni correction:

$$\alpha_{\text{critical}} = 1 - (1 - \alpha_{\text{altered}})^k,$$

where k is the number of comparisons. As a result, all decisions of statistical significance were made at the $\alpha = .001$ level. The confidence intervals (CI) associated with the regression are presented at the 99.9% level.

Additionally, calculations of effect sizes were performed in each regression model using f^2, a variant Cohen's f statistics. The calculations were performed in the following way:

$$f^2 = \frac{R^2}{1 - R^2}.$$

In multiple regression, the effect size is a measure of relation strength between the predictor or a set of predictors and the outcome. The measure of effect size indicates the strength of statistical relationships independent of sample size.

Residuals

In the process of statistical modeling, there is always a sacrifice of precision made for the purpose of real data simplification. The least squares linear regression is by far the most widely used modeling approach.[24] In the least squares method, the unknown parameters (the intercept and slopes) are estimated by minimizing the sum of squared deviations between the real data and the estimates derived from the model. This difference is called the *errors* in the language of parameters and *residuals* in the language of parameter estimates. The following is the statistical definition of errors:

$$\varepsilon_i = Y_i - E(Y_i),$$

where ith error term is defined as the difference between the ith observation and expectation. Thus, the residuals are defined as follows:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i,$$

where ith residual is the difference between the observed value and the value predicted by the model (value of Y on the regression line given X).

In regression, the residuals are assumed to average at zero; this assumption will hold for the overall residuals as well as for residuals segmented by the levels of the variables included in the model. That is, if sex is included in the model as a predictor or covariate, the residuals for only male participants will average at zero; the residuals for only female participants will average at zero as the overall residuals.[25] This assumption, however, need not hold when residuals are segmented by variables that are not included in the regression analysis. In fact, it is well known in the educational literature that when the 1st-year undergraduate college GPA is predicted from the high school GPA and standardized test scores, the overall model produces residuals that average at zero. However, when these residuals are segmented by students' ethnicity, African American and Latino students tend to be over-predicted, while Asian and White students tend to be underpredicted.[6,26] Thus, the evaluation of prediction quality for the variables related to the sample, but not included in the model, may be very informative.

Following this logic, we saved standardized residuals for all 6 regression models and examined them when grouped by the levels of pre–chiropractic GPA, which was not included in the regression models. The standardized residuals are residuals divided by the SD, so that the standardized residuals will have a deviation of 1. The following is the statistical definition of standardized residuals:

$$Z\hat{e}_i = \frac{Y_i - \hat{Y}_i}{\text{SE}(Y_i - \hat{Y}_i)},$$

where $Z\hat{e}_i$ is the ith standardized residual, and SE is the standard error.[27] Since the residuals are standardized, they could be compared across the levels of pre–chiropractic GPA.

RESULTS

Correlations

Table 2 provides 1st-order correlation estimates for the NBCE Part I exam domain scores, demographic characteristics of the sample, pre–chiropractic GPA, and chiropractic GPA. Age revealed negative correlation with all the domains of the Part I exam, which points to younger examinees performing better on the exam. African American and Latino statuses were associated with lower test scores when compared with Other ethnicity status. Asian/Pacific Islander and White statuses, on the other hand, were associated with higher Part I scores. Both pre–chiropractic GPA and GPA obtained in chiropractic programs revealed a positive correlation with the Part I domain scores, which means that students with higher educational attainment were more likely to obtain higher test scores. Yet, educational level acquired prior to enrollment in a chiropractic program did not correlate significantly with the exam scores.

Chiropractic GPA revealed a negative, statistically significant correlation with sex, meaning that female chiropractic students were more likely to have a higher GPA when compared with their male counterparts. Asian examinees were more likely to have higher pre–chiropractic and chiropractic GPA when compared with *Other* ethnicity status. On the other hand, African American examinees were more likely to have lower GPAs. Of interest, education did not correlate with Part I scores; however, usually older students will have more years of education, so the correlation may have been accounted for by age.

Predictive Models

Six regression models (1 for each domain of the Part I) were estimated and tested. Each model tested the predictive validity of the chiropractic GPA after controlling for the effects of demographic characteristics. Standardized residuals were produced and saved after the estimation of each of model.

Table 2 - First-Order Correlations Between Part I Domain Scores, Demographics, Pre–Chiropractic GPA, and Chiropractic GPA

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. General anatomy	—																
2. Spinal anatomy	.750**	—															
3. Physiology	.738**	.713**	—														
4. Chemistry	.692**	.674**	.719**	—													
5. Pathology	.652**	.638**	.650**	.665**	—												
6. Microbiology	.597**	.575**	.624**	.637**	.634**	—											
7. Age	−.163**	−.142**	−.106**	−.136**	−.121**	−.091**	—										
8. Sex	.075*	.080**	.097**	.024	−.055*	−.056*	.084**	—									
9. Asian/Pacific Islander	.205**	.197**	.220**	.162**	.176**	.194**	−.151**	.028	—								
10. African American	−.071**	−.086**	−.125**	−.071**	−.051*	−.060**	.064**	−.035	−.317**	—							
11. Latino	−.127**	−.122**	−.137**	−.096**	−.082**	−.158**	.070**	−.023	−.537**	−.078**	—						
12. Native American	−.121**	−.093**	−.103**	−.066**	−.122**	−.067**	.080**	−.007	−.454**	−.066**	−.111**	—					
13. White	−.009	−.022	.007	−.006	−.028	−.015	.003	.005	−.128**	−.019	−.031	−.027	—				
14. Other	−.025	−.028	−.041	−.051*	−.041	−.057*	.033	.015	−.271**	−.039	−.066**	−.056**	−.016	—			
15. Education	.025	.010	.039	.032	.008	.029	.101**	−.010	−.079**	.124**	.042	.015	−.038	−.027	—		
16. Pre–chiropractic GPA	.423**	.435**	.435**	.386**	.368**	.327**	−.155**	−.018	.089**	−.058**	−.024	−.052*	−.021	.005	.054*	—	
17. Chiropractic GPA	.138**	.109**	.119**	.118**	.097**	.090**	−.103**	−.091**	.077**	−.079**	.010	−.093**	.015	.051*	−.011	.215**	—

* $p < .05$.
** $p < .01$.

In all models, the demographic predictors (age, sex, ethnicity, and education) as a block, revealed statistical significance predicting the scores on the GEA, F_{charge} (7,1834) = 19.43, $p < .001$; SPA, F_{charge} (7,1860) = 16.97, $p < .001$; PHY, F_{charge} (7,1850) = 19.81, $p < .001$; CHE, F_{charge} (7,1835) = 10.56, $p < .001$; PAT, F_{charge} (7,1832) = 12.58, $p < .001$; and MIC, F_{charge} (7,1845) = 13.74, $p < .001$. Based on the models, the block explained between 4% and 8% of variability in domain scores.

After controlling for the effects of demographic covariates, the chiropractic GPA explained 2% of the total variability in GEA scores, F_{charge} (8,1833) = 20.47, $p < .001$; 1% in SPA, F_{charge} (8,1859) = 16.91, $p < .001$; 1% in PHY, F_{charge} (8,1849) = 19.93, $p < .001$; 2% in CHE, F_{charge} (8,1834) = 11.62, $p < .001$; 1% in PAT, F_{charge} (8,1832) = 12.11, $p < .001$; and 1% in MIC, F_{charge} (8,1844) = 13.15, $p < .001$.

The chiropractic GPA was revealed as a significant predictor of all domain scores in the Part I exam. Unstandardized regression coefficients are interpreted as expected change in the dependent variable as a function of 1-unit change in the independent variable.[26] Based on the results, on average, a 28-scaled score-point increase is expected in GEA and SPA scores as a function of 1-unit change (eg, going from C to B or from B to A) in the pre–chiropractic GPA. An approximately 30% increase is expected in PHY, 25 points in CHE, 16 points in PAT, and 18 points in MIC. Regression results are presented in Table 3.

The analyses were not susceptible to multicollinearity as tolerance statistics for predictors ranged from .59 to .97. Effect sizes were calculated using f^2 statistic. According to the Cohen rule of thumb, values ranging from .01 to .14 constitute small effect; values ranging from .15 to .34, medium effect; and values >.34, large effect.[28] The values of effect-size estimates were calculated at each step of the regression (Table 3).

Residuals

One of the assumptions made in regression analysis is that residuals (errors) are independently, identically distributed, following normal distribution with $\mu = 0$ and σ^2 or $e_i \sim N(0, \sigma^2)$, where $i = 1, 2, \ldots, N$.[27] That is, we assume that each residual is sampled from the same normal distribution with a mean of zero and the same variance throughout. Thus, the value of average residual should be zero at every value of X. This assumption was met for all 6 models.

A different perspective was obtained when residuals were split by pre–chiropractic GPA. Examinees who reported their GPA in the range from C to B were assigned to group 1, and those who reported their GPA to be in the range from B to A were assigned to group 2. The average standardized residuals for group 1 (C to B) ranged from −.66 to −.49; and for group 2 (B to A), from .17 to .24. These results show that, on average, lower GPA students tended to be overpredicted (negative average residuals), and higher GPA students tended to be under-predicted (positive average residuals) by as much as a half

Table 3 - Hierarchical Multiple Regression Predicting Part I Domain Scores

	At Step				Final Model				99.9% CI for *B*	
	Δ *R²*	*R²*	Multiple *R*	*f²*	*B*	SE	β	Prob	Lower	Upper
Dependent variable: general anatomy, *n* = 1841										
Step 1: Demographics	.07	.07	.28	.40				<.001		
Age					−2.61	.46	−.13	<.001	−3.79	−1.43
Sex					15.00	3.82	.09	<.001	5.15	24.84
Asian/Pacific Islander					−23.34	10.16	−.08	<.05	−49.55	2.87
African American					−17.01	12.06	−.04	>.05	−48.10	14.08
Latino					−22.91	9.59	−.09	<.05	−47.64	1.81
Native American					−2.90	22.64	.00	>.05	−61.27	55.48
White					14.65	8.15	.08	>.05	−6.38	35.67
Education					7.66	2.24	.08	<.001	1.89	13.43
Step 2: GPA	.02	.09	.30	.03				<.001		
Chiropractic GPA					28.29	5.75	.11	<.001	13.47	43.10
Dependent variable: spinal anatomy, *n* = 1867										
Step 1: Demographics	.06	.06	.25	.34				<.001		
Age					−2.85	.58	−.11	<.001	−4.34	−1.36
Sex					18.87	4.81	.09	<.001	6.46	31.28
Asian/Pacific Islander					−17.62	12.86	−.05	>.05	−50.77	15.53
African American					−27.33	15.19	−.05	>.05	−66.51	11.84
Latino					−23.98	12.12	−.07	<.05	−55.24	7.28
Native American					−15.47	27.93	−.01	>.05	−87.49	56.55
White					21.32	10.32	.09	<.05	−5.29	47.93
Education					7.81	2.81	.06	<.001	.57	15.05
Step 2: GPA	.01	.07	.27	.03				<.001		
Chiropractic GPA					27.62	7.27	.09	<.001	8.88	46.36
Dependent variable: physiology, *n* = 1857										
Step 1: Demographics	.08	.08	.28	.40				<.001		
Age					−1.86	.57	−.07	<.001	−3.33	−.39
Sex					21.63	4.79	.10	<.001	9.28	33.99
Asian/Pacific Islander					−26.09	12.62	−.07	<.05	−58.61	6.44
African American					−52.91	15.02	−.10	<.001	−91.64	−14.17
Latino					−34.36	11.94	−.10	<.01	−65.14	−3.58
Native American					16.14	28.51	.01	>.05	−57.36	89.65
White					19.20	10.07	.08	>.05	−6.78	45.17
Education					12.23	2.79	.10	<.001	5.04	19.42
Step 2: GPA	.01	.09	.30	.03				<.001		
Chiropractic GPA					30.24	7.21	.09	<.001	11.65	48.84
Dependent variable: chemistry, *n* = 1842										
Step 1: Demographics	.04	.04	.22	.29				<.001		
Age					−2.19	.48	−.11	>.05	−3.43	−.94
Sex					6.21	4.01	.04	>.05	−4.14	16.55
Asian/Pacific Islander					−2.03	10.64	−.01	>.05	−29.46	25.40
African American					−12.97	12.64	−.03	>.05	−45.56	19.63
Latino					−9.97	10.05	−.04	>.05	−35.90	15.95
Native American					8.74	23.80	.01	>.05	−52.63	70.10
White					20.54	8.50	.11	<.05	−1.39	42.47
Education					8.83	2.35	.09	<.001	2.77	14.90
Step 2: GPA	.02	.06	.24	.03				<.001		
Chiropractic GPA					24.94	6.04	.10	<.001	9.36	40.52
Dependent variable: pathology, *n* = 1839										
Step 1: Demographics	.04	.04	.22	.29				<.001		
Age					−1.74	.46	−.09	<.001	−2.93	−.54
Sex					−7.94	3.88	−.05	<.05	−17.93	2.06
Asian/Pacific Islander					−18.66	10.31	−.06	>.05	−45.25	7.93
African American					−6.17	12.32	−.01	>.05	−37.94	25.60
Latino					−6.80	9.76	−.03	>.05	−31.96	18.37
Native American					−12.91	22.97	−.01	>.05	−72.15	46.33

Table 3 - Continued.

	At Step				Final Model				99.9% CI for *B*	
	Δ *R²*	*R²*	Multiple *R*	*f²*	*B*	SE	β	Prob	Lower	Upper
White					20.83	8.27	.12	<.05	−.50	42.16
Education					5.12	2.27	.05	<.05	−.74	10.99
Step 2: GPA	.01	.05	.24	.03				<.001		
Chiropractic GPA					16.44	5.86	.07	<.001	1.33	31.55
Dependent variable: microbiology, *n* = 1852										
Step 1: Demographics	.05	.05	.23	.31				<.001		
Age					−1.25	.50	−.06	<.05	−2.54	.04
Sex					−10.09	4.21	−.05	<.05	−20.95	.78
Asian/Pacific Islander					−9.68	11.12	−.03	>.05	−38.35	18.98
African American					−17.55	13.29	−.04	>.05	−51.81	16.71
Latino					−32.27	10.46	−.11	<.01	−59.25	−5.29
Native American					−5.88	25.04	−.01	>.05	−70.45	58.69
White					19.05	8.88	.10	<.05	−3.84	41.94
Education					7.04	2.47	.06	<.01	.67	13.40
Step 2: GPA	.01	.06	.25	.03				<.001		
Chiropractic GPA					17.76	6.36	.06	<.001	1.37	34.15

Prob = probability.

the SD. Averages for residuals split by the pre–chiropractic GPA are presented in Table 4.

Most students reported their pre–chiropractic GPA ranged from B to A; for them, the predicted exam score is lower than the actual score. For students in the C to B range, the actual obtained score is lower than predicted, which results in a negative residual value.

Pass/Fail Rates

In the following step, we divided the sample into 2 groups according to examinees' pre–chiropractic and chiropractic GPAs. Those with a pre–chiropractic GPA of 3.5 or higher were classified as *very good students,* and those with a pre–chiropractic GPA lower than 3.5 were classified as *other students*. We performed the same classification based on examinees' chiropractic GPA.

The differences in the pass/fail rates varied as a function of pre–chiropractic GPA. The pass rates for *very good students* were 97.2% and 98.3% for the Part I and Part II

Table 4 - Standardized Residuals by Pre–Chiropractic GPA

GPA Range	Model Predicting	*n*	Min	Max	Mean	SD
C to B	General anatomy	481	−3.41	1.72	−.62	.81
	Spinal anatomy	496	−3.10	1.43	−.65	.81
	Physiology	493	−2.58	1.40	−.66	.80
	Chemistry	482	−2.90	2.26	−.57	.85
	Pathology	479	−3.91	1.64	−.55	.82
	Microbiology	486	−2.69	2.16	−.49	.89
B to A	General anatomy	1360	−2.75	3.33	.22	.96
	Spinal anatomy	1371	−2.97	3.08	.24	.95
	Physiology	1364	−3.48	2.77	.24	.95
	Chemistry	1360	−2.92	3.29	.20	.97
	Pathology	1360	−2.47	3.50	.19	.98
	Microbiology	1366	−2.99	3.16	.17	.97

exams, respectively (Fig. 1). The rates for *other students* were 65.3% and 61.3% for the Part I and Part II exams, respectively. The differences in pass rates were also discovered as a function of chiropractic GPA. The pass rates for *very good students* were 82.7% and 74% for the Part I and II exams, respectively. The pass rate for *other students* was 71% for the Part I and II exams (Fig. 2).

DISCUSSION

The goal of this paper was to bring the attention of the chiropractic educational community to the validity issue of utilizing test scores outside of the primary use. In addition, we aimed at building a validity argument for the connection between the chiropractic GPA and standardized test scores, particularly, for the purposes of educational accountability. We argue that the degree and depth of standardized test scores used for educational accountability in chiropractic colleges should be well supported by the evidence of validity.

Camara et al[29] stated that "additional uses of assessment scores would unavoidably generate additional claims, each of which requires the same type of interpretative argument and accumulation of evidence as the original use (p. 15)." Therefore, examining the predictive validity of the chiropractic GPA for the Part I scores is the first step in the direction of collecting evidence to build this validity argument.
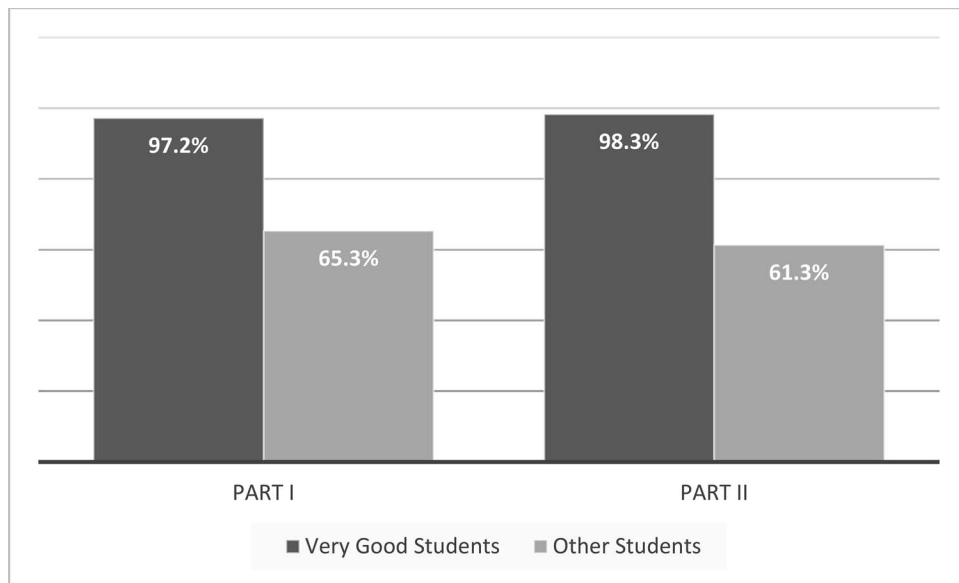
This study builds on the cascading evidence from educational literature by providing additional results suggesting that undergraduate GPA as well as the GPA obtained in doctor of chiropractic programs are predictive of future performance. Previous research has intensively examined the predictive validity of GPA in combination with standardized test scores.[6,30]

The uniqueness of this study is in the manner that standardized test scores were included in the analyses.
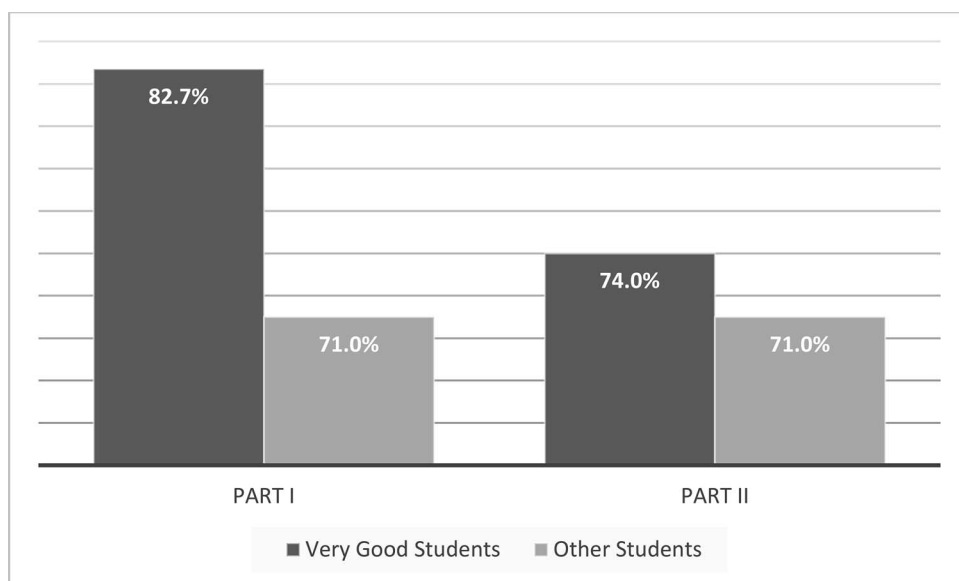
Figure 1 - Part I and Part II pass rates disaggregated by pre–chiropractic GPA.

Instead of treating the scores as the predictor of future performance, the Part I scores were predicted by the models comprising the GPA. Therefore, this work becomes pioneering in the chiropractic profession as the results provide a first glance at the connection between the requirements for admission to chiropractic training and the ability to pass the exams required to enter the profession.

This study demonstrated that the results obtained from the chiropractic profession are in line with the results of similar studies conducted in mainstream education and education in other professions. The GPA, being the strongest predictor of future success, was confirmed as a predictor of the success in the chiropractic profession. The

correlations between the domains of the Part I and pre–chiropractic GPA were moderately strong. The chiropractic GPA, however, was less related to the scores on the Part I exam. After controlling for the demographic predictors, the effect sizes associated with the inclusion of GPA in the models were consistently small. This may be explained by possible grade inflation in chiropractic institutions and probable variability in educational quality among chiropractic institutions. Educators can use the results of this study to identify students with potential problems early on. Considering that modeling was done at a domain level, the results may be used to determine which preparation tools may be valuable to help students to pass exams.

Figure 2 - Part I and Part II pass rates disaggregated by chiropractic GPA.

While, to some extent, this study confirmed the predictive validity of the chiropractic GPA for the Part I exam scores, the results identify intriguing differences in the prediction patterns when data were disaggregated by the students' merit. The exam scores for *very good students* were underpredicted, while the scores for *other students* were overpredicted.

According to Zwick,[31] Anastasi[32] and Einhorn and Bass[33] were first to point out that "two groups can have identical regression equations but unequal residual variances. In this situation, members of 2 groups will have the same predicted criterion score for all values of the predictor but will not have the same probability of exceeding a cut point on the criterion."[31] (p. 35)

The possibility of under- and overprediction is a well-known phenomenon to psychometricians and educational researchers. Mattern et al[34] documented differential prediction of the 1st-year college GPA when regressed on the high school GPA and SAT scores. The standardized residuals were disaggregated and averaged by sex and ethnicity of test takers. Their results revealed underprediction for females and overprediction for males. In another series of studies, patterns of overprediction of the 1st-year college GPA were found for African American and Latino students, while White and Asian students were underpredicted.[6,35]

Numerous undercurrents may lead to the patterns of the statistical phenomenon documented by this study including the chance alone. Another explanation is offered by researchers who studied the effects of measurement error on group differences in modeling academic achievement. They recognized that the unreliability of predictors influences the degree of bias in least squares regression.[36,37] Errors of measurement may arise from different sources of variability in score, while the measured scores themselves are assumed to be on the same scale. For example, if 2 measurements of the same magnitude are assumed to be on the same scale, they are expected to be equal.[38] The inconsistencies of the measure result in a systematic error, which in turn affects the validity of predictors.

In the framework of this study, students may have different degrees of measurement error associated with their chiropractic GPA. The systematic and erroneous differences in the quality of prediction observed in this study may be a direct function of chiropractic GPA's unreliability. The reliability and validity of GPA across chiropractic colleges is unknown. However, Bacon and Bean[39] examined the psychometric qualities of GPA using grade and admission data extracted from a database of a university's office of institutional research. Their results demonstrated a very high reliability for overall 4-year GPA, and a much lesser reliability for GPA calculated within major.

Limitations

Recently, criticism of predictive models encompassing GPA and standardized test scores has emerged in the educational literature. The argument is built against the correlational evidence of GPA's predictive superiority presented by the regression models. Because of possible violations of statistical assumptions made by linear models, the correlation coefficient becomes less than ideal for prediction; the use of standardized test scores is advised instead.[40] In this study, we used linear models, which assume normal distribution of the outcome variable—an assumption upheld after our diagnostic analysis.

The Part I scores were specified as the outcome in the study, which constitutes a further limitation. There were 2 reasons why this decision was made: (1) we used the data available at the time of study, and (2) our experience shows that pre–chiropractic GPA, which is included as predictor, is more related to the test of basic sciences. Future studies may benefit from including scores from other exams as model outcomes.

Methodologically, our study's design involved a nonexperimental approach evaluating cross-sectional variables. Thus, causal relationships may not be established between the predictors and the outcomes. Additionally, except for the Part I scores, all measures included in this study were self-reported by the examinees who may have engaged in socially desirable responding. For instance, we cannot rule out the possibility that participants overstated their GPA or level of education preceding admission to a chiropractic college.

The sample size used in this study, although adequate for the performed analytics, is relatively small. The samples further varied by size when residuals were disaggregated by the pre–chiropractic GPA; therefore, the degree to which our findings are generalizable is unknown. Yet, we hope that this study is informative and motivating to conduct additional research, which will benefit and advance the chiropractic profession.

CONCLUSIONS

Statistically, both pre–chiropractic GPA and chiropractic GPA are connected to the domain scores of the Part I exam. Yet, after controlling for the variability in scores explained by the demographic characteristics of the sample, the addition of the chiropractic GPA to the regression models serves very little for the quality of prediction. The disconnect may have several possible explanations including the inflation of the chiropractic GPA, the quality of training in basic sciences before the admission to a chiropractic college and/or during a chiropractic program, the misalignment between the chiropractic curricula and the Part I exam, and simply error in measurement. This issue calls for a deeper investigation. To rule out the misalignment of the curricula and the exam, the NBCE will conduct a Delphi study to reaffirm the content validity of Part I and Part II exams.

Additionally, the patterns of prediction in models predicting Part I exam scores from the chiropractic GPA are systematically different for students with higher pre–chiropractic GPA and *other students*. This is an alarming dynamic. The NBCE and chiropractic institutions need to take a stand against standardized test score misuses and advocate together for fair assessment and quality educa-

tion, while building the validity case for each interpretation of assessment scores.

FUNDING AND CONFLICTS OF INTEREST

About the Authors

Igor Himelfarb (corresponding author) is the director of psychometrics and research at the NBCE (901 54th Ave, Greeley, CO 80634; ihimelfarb@nbce.org). Bruce Shotts is the director of written examinations at the NBCE (901 54th Ave, Greeley, CO 80634; bshotts@nbce.org). Andrew Gow is the director of practical examinations at the NBCE (901 54th Ave, Greeley, CO 80634; agow@nbce.org). This article was received March 22, 2020; revised August 21, 2020, December 7, 2020, April 29, 2021, and June 3, 2021; and accepted June 10, 2021.

REFERENCES

1. Braun HI. *Using Student Progress To Evaluate Teachers: A Primer on Value-Added Models.* Policy Information Perspective. Princeton, NJ: Educational Testing Service; 2005. Report No. 730194.
2. Messick S. Test validity and the ethics of assessment. *Am Psychol.* 1980;35(11):1012–1027.
3. Crocker L, Algina J. *Introduction to Classical and Modern Test Theory.* New York, NY: Harcourt Brace Jovanovich College Publishers; 1986.
4. Cronbach LJ. Five perspectives on the validity argument. In: Wainer H, Braun HI, eds. *Test Validity.* Hillsdale, MI: Lawrence Erlbaum Associates; 1988:3–17.
5. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing.* Washington, DC: AERA, APA, NCME; 2014.
6. Zwick R, Himelfarb I. The effect of high school socioeconomic status on the predictive validity of SAT scores and high school grade-point average. *J Educ Meas.* 2011;48(2):101–121.
7. Geiser S, Studley R. UC and the SAT: predictive validity and differential impact of the SAT and SAT II at the University of California. In: Zwick R, ed. *Rethinking the SAT: The Future of Standardized Testing in University Admissions.* New York, NY: RoutledgeFalmer; 2004:125–153.
8. Sarbin TR. A contribution to the study of actuarial and individual methods of prediction. *Am J Sociol.* 1943;48:593–602.
9. Zwick R. *Disentangling the Role of High School Grades, SAT Scores, and SES in Predicting College Achievement.* Princeton, NJ: Educational Testing Service; 2013. Report No. ETS RR–13-09.
10. Dobry AM, Murphy PD, Schmidt DM. Predicting teacher competence. *Action Teach Educ.* 1985;7(1–2):69–74.
11. Nelson B, Wood L. The competency dilemma. *Action Teach Educ.* 1985;7(1–2):45–58.
12. Olstad B, Beal JL, Foster CD, Marrett AV. *A Validity Study of NTE General Knowledge Component as a Predictive. Instrument for Successful Student Teaching.* Seattle, WA: Teacher Education Research Center, University of Washington; 1985. Report No. 85-1.
13. Mehrens WA, Phillips SE. Using college GPA and test scores in teacher licensure decisions: conjunctive versus compensatory models. *Appl Meas in Educ.* 1989;2(4):277–288.
14. Camara WJ, Echternacht G. *The SAT I and High School Grades: Utility in Predicting Success in College.* New York, NY: The College Board; 2000. Report No. RN-10.
15. Hu NB. Measuring the weight of high school GPA and SAT scores with second term GPA to determine admissions/financial aid index: a case study. Paper presented at: 42nd Annual AIR Forum; June 2–5, 2002; Toronto, Canada.
16. Noble J, Sawyer R. Accuracy of high school grades and college admissions test scores for predicting different levels of academic achievement in college. Paper presented at: 42nd Annual AIR Forum; June 2–5, 2002; Toronto, Canada.
17. Esparza SD. *Predictors of Success on Professional Credentialing Examinations of Athletic Training Undergraduates.* Minneapolis, MN: Walden University; 2012.
18. Arathuzik D, Aber C. Factors associated with National Council Licensure Examination-registered nurse success. *J Prof Nurs.* 1998;14(2):119–126.
19. Hu Y, Martindale JR, LeGallo RD, White CB, McGahren ED, Schroen AT. Relationships between preclinical course grades and standardized exam performance. *Adv Health Sci Educ.* 2016;21(2):389–399.
20. Efurd MG. *Predictors of Academic Success for the National Board Dental Hygiene Examination and the Southern Regional Testing Agency Clinical Exam* [dissertation]. Fayetteville, AR: University of Arkansas; 2012.
21. Christensen MG, Kollash MW, Hyland JK. Practice Analysis of Chiropractic. Greeley, CO: National Board of Chiropractic Examiners; 2015.
22. Hair JF Jr, Hult GT, Ringle C, Sarstedt M. *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM).* Los Angeles, CA: Sage publications; 2016.

23. Bonferroni C. *Teoria statistica delle classi e calcolo delle probabilità*. Firenze: Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze; 1936.

24. Wolberg J. *Data Analysis Using the Method of Least Squares: Extracting the Most Information from Experiments*. Berlin: Springer; 2005.

25. Faraway J. *Linear Model with R*. 2nd ed. Boca Raton, FL: CRC Press; 2014.

26. Clearly TA. Test bias: prediction of grades of negro and white students in integrated colleges. *J Educ Meas*. 1968;5:115–124.

27. Kutner MH, Nechtsheim C, Neter J, Li W. *Applied Linear Statistical Models*. 5th ed. Boston, MA: McGraw-Hill; 2005.

28. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Academic Press; 2013.

29. Camara WJ, Mattern K, Croft M, Vispoel S, Nichols P. A validity argument in support of the use of college admissions test scores for federal accountability. *Educ Meas Issues Pract*. 2019;38(4):12–26.

30. Zwick R. *Who Gets in? Strategies for Fair and Effective College Admissions*. Cambridge, MA: Harvard University Press; 2017.

31. Zwick R. Fairness in measurement and selection: statistical, philosophical, and public perspectives. *Educ Meas Issues Pract*. 2019;38(4):34–41.

32. Anastasi A. *Psychological Testing*. 6th ed. New York, NY: Macmillan; 1988.

33. Einhorn H, Bass A. Methodological considerations relevant to discrimination in employment testing. *Psych Bull*. 1971;75:261–269.

34. Mattern KD, Shaw EJ, Williams FE. *Examining the Relationship Between the SAT, High School Measures of Academic Performance, and Socioeconomic Status: Turning Our Attention to the Unit of Analysis*. New York, NY: The College Board; 2008. Report No. RN-36.

35. Zwick R, Green J. New perspectives on the correlation of SAT scores, high school grades, and socioeconomic factors. *J Educ Meas*. 2007;44:23–45.

36. Wolfle LM, Robertshaw D. *Racial Differences in Measurement Error in Educational Achievement Models*. New York, NY: American Educational Research Association; 1982.

37. Culpepper SA. Evaluating EIV, OLS, and SEM estimators of group slope differences in the presence of measurement error: the single-indicator case. *App Psych Meas*. 2012;36(5):349–374.

38. Kane M. *Errors of Measurement, Theory, and Public Policy*. Princeton, NJ: Educational Testing Service, Policy Evaluation and Research Division; 2010.

39. Bacon RB, Bean B. GPA in research studies: an invaluable but neglected opportunity. *J Mark Educ*. 2006;28(1):35–42.

40. Sawyer R. Beyond correlations: usefulness of high school GPA and test scores in making college admissions decisions. *App Meas Educ*. 2013;26:89–112.