The Predictive Validity of the MCAT for Medical School Performance and Medical Board Licensing Examinations: A Meta-Analysis of the Published Research

Tyrone Donnon, PhD, Elizabeth Oddone Paolucci, PhD, and Claudio Violato, PhD

Abstract

Purpose
To conduct a meta-analysis of published studies to determine the predictive validity of the MCAT on medical school performance and medical board licensing examinations.

Method
The authors included all peer-reviewed published studies reporting empirical data on the relationship between MCAT scores and medical school performance or medical board licensing exam measures. Moderator variables, participant characteristics, and medical school performance/medical board licensing exam measures were extracted and reviewed separately by three reviewers using a standardized protocol.

Results
Medical school performance measures from 11 studies and medical board licensing examinations from 18 studies, for a total of 23 studies, were selected. A random-effects model meta-analysis of weighted effects sizes (*r*) resulted in (1) a predictive validity coefficient for the MCAT in the preclinical years of $r = 0.39$ (95% confidence interval [CI], 0.21–0.54) and on the USMLE Step 1 of $r = 0.60$ (95% CI, 0.50–0.67); and (2) the biological sciences subtest as the best predictor of medical school performance in the preclinical years ($r = 0.32$ 95% CI, 0.21–0.42) and on the USMLE Step 1 ($r = 0.48$ 95% CI, 0.41–0.54).

Conclusions
The predictive validity of the MCAT ranges from small to medium for both medical school performance and medical board licensing exam measures. The medical profession is challenged to develop screening and selection criteria with improved validity that can supplement the MCAT as an important criterion for admission to medical schools.

Acad Med. 2007; 82:100–106.

The MCAT continues to be widely used for screening and selection for many medical schools in the United States and Canada.[1] But how good is the MCAT at predicting students' performance in medical school and beyond? Notwithstanding substantial research efforts, the predictive validity of the MCAT and, in particular, its subtest domains remains unclear. Specifically, Baker et al[2] computed a range of predictive validity coefficients from $r = -0.18$ to 0.13 on the MCAT subtests in a small sample of 63 students on first- and second-year medical school performance measures, and Hojat et al.[3] found that for 1,271 and 1,006 students, the writing sample subtest had a coefficient of zero on both the USMLE Step 1 ($r = 0.02$) and Step 2 ($r = 0.04$) exams. Conversely, Swanson[4] derived a range of MCAT subtest coefficients ($r = 0.14$ to 0.52) on the USMLE Step 1 examination based on a large sample of 11,145 students, and in a study of 27,406 students, Julian[5] found that their total MCAT scores correlated moderately well across all three USMLE Step examinations (Step 1, $r = 0.61$; Step 2, $r = 0.49$; Step 3, $r = 0.49$). Accordingly, the major purpose of the present study was to conduct a meta-analysis of the predictive validity of the MCAT and its various subtests on medical school and licensing examination performance measures, to determine both the magnitude of the coefficients as well as their confidence intervals.

In part because the MCAT has evolved over the course of a number of years and in part because of research challenges, the predictive validity of this test within and beyond medical school needs further exploration. Moreover, the MCAT remains a high-stakes examination and is widely used for medical school admission as a selection criterion into the medical profession. The main purpose of our study, therefore, was to conduct an empirical integration of all published data—a meta-analysis—of the predictive validity of the post-1991 version of the MCAT. We focused on two specific questions: What are the magnitude and the confidence intervals of the predictive validity coefficients of the total MCAT and its subtests on (1) medical school performance and (2) medical board licensing examinations? To address these two questions, we performed a systematic review and empirical integration of published research on the predictive validity of the current version of the MCAT.

Dr. Donnon is a faculty member, Medical Education and Research Unit, and assistant professor, Department of Community Health Sciences, Faculty of Medicine, University of Calgary, Calgary, Canada.

Dr. Oddone Paolucci is research associate, Applied Psychology, Faculty of Education, University of Calgary, Calgary, Canada.

Dr. Violato is director, Medical Education and Research Unit, and professor, Department of Community Health Sciences, Faculty of Medicine, University of Calgary, Calgary, Canada.

Correspondence should be addressed to Dr. Donnon, Medical Education and Research Unit, G705, Undergraduate Medical Education, Faculty of Medicine, University of Calgary, 3330 Hospital Drive NW, Calgary, AB Canada, T2N 4N1; telephone: (403) 210-9682; fax: (403) 270-2681; e-mail: (tldonnon@ucalgary.ca).

Method

Selection of studies

For this study, we followed the guidelines for the reporting of observational studies in a meta-analysis.[6] In addition to a MEDLINE (January 1991 to October 2005) search, the PsychINFO (January 1991 to October 2005) and ERIC

(January 1991 to October 2005) databases were also searched. Because the MCAT is used primarily by American and Canadian medical schools, we restricted our search to English-language publications. To be included, a study had to meet the following criteria: (1) used the current version of the MCAT or its subtests (the independent variables) that was introduced in 1991, (2) presented empirical findings of MCAT scores related to at least one medical school performance or medical board licensing dependent variable, (3) employed psychometrically sound dependent measures (i.e., standardized instruments, summative examinations, objectively scored observational ratings, etc.,), and (4) was published in print form in a refereed, peer-reviewed journal. Restricting the inclusion to articles that were published only in refereed journals enhanced study quality in that only research that had been peer reviewed was used.

Data extraction

Our initial search yielded 61 journal articles; 23 studies met the inclusion criteria requirements, and 38 articles failed to meet all four inclusion criteria (e.g., not a referred journal[7] or no empirical data[1]. We developed a coding protocol that included each study's title, author(s)' name(s), year, source of publication, country of origin, study design (predictive or comparison study), measures of MCAT, types of medical school assessment (multiple choice, objective structured clinical examinations, standardized tests), and medical board licensing performance examination (i.e., USMLE Step 1, Step 2, Step 3, or comparable licensure examination). When information for the following moderator variables was available, we also coded for the following: sex, race, ethnicity, underrepresented minority, type of medical school curriculum (e.g., systems based, problem-based learning (PBL)), and location of medical school. All 23 articles were independently coded by two coders (TD and EOP), and any discrepancies (e.g., potential multiple publication) were reviewed by a third coder (CV). After discussions among the coders and iterative reviews, we achieved 100% agreement on data coding by all three coders.

Statistical analysis

The statistical analysis was performed using the Comprehensive Meta Analysis software program (version 1.0.23, Biostat Inc., Englewood, NJ). We used the correlation coefficient (Pearson's product–moment, r) as the effect size. We selected MCAT total or subtests as the independent variables and medical school performance or medical board licensing exams as the dependent variables.[8] For 17 (73.9%) of 23 studies identified, a correlation between the MCAT scores and medical performance outcomes was provided in the results section, allowing for easy data extraction. For the remaining six studies, we calculated r from other reported data. Thus, correlations were derived through P values in two studies (8.7%), through beta weights in two studies (8.7%), and, in two separate studies, bivariate r^2 values and F ratios, respectively, using the standard conversions of these statistics to effect sizes.[8] We plotted the standard deviations of the unweighted effect sizes to examine the heterogeneity of studies and to identify outliers and naturally occurring groupings that could be explained by examining moderators.

We employed a random-effects model (DerSimonian and Laird) in combining the weighted and unweighted effect sizes, because this model reflects a conservative estimate of the between-study variance of both medical school performance and licensing board examination outcomes.[9] One of the main concerns in the integration of results from different studies is the diversity that can occur in the research designs and methods those studies used. Although a fixed-effects model assumes that the testing of students' performance between medical schools is the same, the random-effects model incorporates the more common effects of heterogeneity in the analysis. In addition, Forrest plots with Cochran Q tests for heterogeneity of effect sizes were done.[10] Based on the assumption of a null hypothesis, however, the absence of a significant P value for Q does not by itself imply homogeneity because it could reflect low power within studies rather than actual consistency. Therefore, a review of the actual dispersion of the studies on the Forrest plot becomes an important visual indicator for consistency between studies.

The interpretation of the magnitude of the effect size for linear correlations is based on Cohen's[11] suggestions that an r of 0.10 ($r^2 = 0.01$) be considered "small," an r of 0.30 ($r^2 = 0.09$) as "medium," and an r of 0.50 ($r^2 = 0.25$) as "large." To assess for publication bias, we applied Rosenthal's "file drawer" method to determine the number of unpublished studies with a mean observed effect of zero that would be needed to make the noted effect no longer significant.[12] In an extreme view of the file drawer problem, journals publish the 5% of studies that show Type I errors, whereas 95% of the studies that show nonsignificant results (i.e., $P > .05$) remain unpublished, in the file drawers of researchers' offices. The approach in dealing with the file drawer problem is to calculate the number of null results studies that would be needed to bring the probability of a Type I error to a desirable level of significance (i.e., $P = .05$). Finally, because only top performers on the MCAT are generally admitted to medical school, this produces an underestimate of r because of this restriction of range. This is a well-known phenomenon, for example, when the range for one variable is restricted due to selection processes. Therefore, we performed adjustments for restriction of range of the MCAT and its subtests.

Results

The 23 studies included in the present study are shown in Table 1 and list the reported MCAT subtests and medical school performance and medical board licensing exam measures with corresponding nonstandardized, unweighted effect sizes. The table includes a list of contrasting subgroup samples for (1) three studies that identified underrepresented minority (URM) medical students, (2) one study that contrasted a traditional program with a PBL program, and (3) one study that contrasted an older and a current version of the MCAT. The sample size of the studies range from a single cohort year of 25 students from one medical school[13] to 22,495 medical students from a total of 112 medical schools that had completed the USMLE Step 1 examination for the first time in either June of 1993 or June of 1994.[4] Information on specific demographic characteristics such as students' sex, age, or socioeconomic status was very rarely

Table 1

Characteristics of 23 Studies of the MCAT with Medical School Performance and/or Medical Licensing Examination Measures

Study source	Sample size	Contrast	MCAT measure†	Outcome‡	r_{UWM}§
Baker et al,[2] 2000	63		VR, BS, PS, WS	MSP-Y1	0.13, 0.30, 0.06, −0.18
	63		VR, BS, PS, WS	MSP-Y2	0.13, 0.02, 0.06, −0.08
	63		VR, BS, PS, WS	MBL-S1 (COMLEX-USA1)	0.12, 0.26, 0.22, −0.13
Basco et al,[23] 2002	933		VR, BS, PS	MBL-S1	0.397, 0.553, 0.491
Dixon,[24] 2004	174		VR, BS, PS	MSP-Y1	0.16, 0.40, 0.34
	174		VR, BS, PS	MSP-Y2	0.17, 0.26, 0.18
	174		VR, BS, PS	MBL-S1 (COMLEX-USA1)	0.16, 0.44, 0.43
	174		VR, BS, PS	MBL-S2 (COMLEX-USA2)	0.31, 0.34, 0.34
Edelin and Ugbolue,[13] 2001	25	URM	T4	MSP-Y1, MSP-Y2	0.00, 0.00
	14	URM	T4	MBL-S1	0.53
Evans et al,[25] 2003	75		T4	MBL-S2 (COMLEX-USA2)	0.41
Gilbert et al,[26] 2002	355		VR, BS, PS, WS	MBL-S1	0.34, 0.57, 0.49, 0.10
	355		VR, BS, PS, WS	MBL-S2	0.35, 0.43, 0.35, 0.11
Giordani et al,[27] 2001	443	Traditional	VR, BS, PS	MSP-Y1	0.26, 0.43, 0.42
	58	PB program	VR, BS, PS	MSP-Y1	0.06, 0.42, 0.32
	15	PB/PMF	VR, BS, PS	MSP-Y1	−0.24, 0.11, −0.17
Haist et al,[28] 2003	275		VR, BS, PS, WS	MSP-Y4	0.14, 0.12, 0.06, 0.07
Hojat et al,[3] 2000	1,271		WS	MBL-S1	0.020
	1,006		WS	MBL-S2	0.039
Huff et al,[29] 1999	1,968		T4	MSP-Y3	0.32
Julian,[5] 2005	*4,076		T4	MSP-PC, MSP-Y3	0.44, .32
	27,406		T4	MBL-S1	0.61
	26,752		T4	MBL-S2	0.49
	25,170		T4	MBL-S3	0.49
Kasuya et al,[30] 2003	258	URM	VR, BS, PS, WS, T4	MBL-S1	0.219, 0.548, 0.574, 0.045, 0.543
	258	URM	VR, BS, PS, WS, T4	MBL-S2	0.274, 0.344, 0.310, 0.122, 0.410
Kossoff et al,[31] 1999	400		T4	MSP-PC	0.113
Kulatunga-Moruzi and Norman,[32] 2002	52		VR, T4	MBL-S2 (LMCC Part I)	0.32, 0.33
	44		VR, T4	MBL-S3 (LMCC Part II)	0.17, 0.07
Mitchell et al,[33] 1994	*1,512		T4	MSP-Y1	0.58
Ogunyemi and Taylor-Harris,[34] 2004	171	URM	T4	MSP-Y3 (Ob/Gyn exam)	0.48
	171	URM	T4	MBL-S1	0.66
Peterson and Tucker,[35] 2005	285		VR, BS, PS, WS	MBL-S1	0.27, 0.48, 0.38, 0.13
Simon et al,[36] 2002	355		VR, BS, PS, WS	MBL-S1	0.35, 0.48, 0.44, 0.10
Spellacy,[37] 1998	165		T4	MBL-S2 (Ob/Gyn Exam)	0.22
Swanson et al,[4] 1996	*11,350	MCAT3			
	*11,145	MCAT4	VR, BS, PS, WS	MBL-S1	0.33, 0.52, 0.49, 0.14
Veloski,[38] 2000	1,940	USMLE	VR, BS	MBL-S1	0.13, 0.34
	1,660	USMLE	VR, BS	MBL-S2	0.21, 0.22
	650	USMLE	VR, BS	MBL-S3	0.28, 0.11
Violato and Donnon,[39] 2005	597		VR, BS, PS, WS	MBL-S2 (LMCC Part I)	0.26, 0.19, −0.03, −0.03
Wiley and Koenig,[40] 1996	*1,764		T4	MSP-Y1, MSP-Y2, MSP-PCMBL-S1	0.67, 0.62, 0.640.72

* Indicates sample size was estimated based on a calculation of 126 students (median) for each medical school institute included in the study analysis.

† T4 = MCAT total test score. MCAT subtests: VR = verbal reasoning, BS = biological sciences, PS = physical sciences, WS = writing sample. URM stands for underrepresented minorities.

‡ Medical school performance (MSP) Outcomes include: MSP-Y1 = year 1 preclinical, MSP-Y2 = year 2 preclinical, MSP-Y3 = year 3 clinical, MSP-Y4 = year 4 clinical, MSP-PC = preclinical years, MSP-CS = clinical years, MSP-T = total. Medical board licensing (MBL) examination outcomes include: MBL-S1 = USMLE Step 1, MBL-S2 = USMLE Step 2, MBL-S3 = USMLE Step 3. Comparable USMLE shelf examinations are identified as being equivalency measures for potential performance on the similar full version of the USMLE examinations. The Medical Council of Canada provides similar licensure examinations and are identified correspondingly as LMCC Part I and Part II.

§ r_{UWM} refers to the association as defined by Pearson's product–moment correlation unweighted mean.

reported or even referred to in the identified studies.

As shown in Table 1, the independent measure may consist of the MCAT total score, a single MCAT subtest, several of the subtests, or a combination thereof. Although contrasts between moderator variables were generally not reported, results on URM performance appear in three of the primary studies included in the meta-analysis. In each study, the unweighted mean Pearson product–moment correlation (r_{UWM}) is provided between the MCAT score and the corresponding medical school performance and/or medical board licensing exam measures.

Medical school basic science and clinical performance measures

Of the 11 studies that reported data on medical school performance measures, eight (72.7%) had results for basic sciences or preclinical performance in medical school, and four (36.4%) for clerkship or clinical years (Table 2). Students' medical school performance in the basic sciences or preclinical years was correlated with the biological sciences subtest at $r = 0.32$ (95% CI, 0.21–0.42), the physical sciences subtest at $r = 0.23$ (95% CI, 0.09–0.36), and the verbal reasoning subtest at $r = 0.19$ (95% CI, 0.12–0.25). As shown in Table 2, we calculated restriction of range adjustments for all of the MCAT subtests (except for the writing sample) and showed a slight increase in the coefficients to $r = 0.40$, $r = 0.26$, and $r = 0.24$, respectively.[14] Two separate studies indicated that $r = 0.0$ for the writing sample subtest with medical school performance for both the basic sciences ($r = -0.13$; 95% CI, −0.30 to 0.05) and clinical ($r = 0.07$; 95% CI, −0.05 to 0.19) dependent variables. The number of file drawer or unreported studies with a mean observed effect of zero required to bring the new overall P to the level just below significance of $P = .5$ is not applicable for the writing sample and physical sciences subtests in the clinical years, because the 95% confidence intervals indicate nonsignificance already. As shown in Table 2, the number of unpublished studies with a mean observed effect of zero that would be needed to make the effect size no longer significant varies from a single to as many as 46 similar studies.

Medical board licensing examination measures

Of the 19 studies that reported data for medical board licensing examination scores, the majority (16; 83.3%) included either the USMLE Step 1, Step 2, or Step 3 examinations. In four (21.1%) of the studies, standardized specialty examinations were used as the dependent measure and were treated as corresponding Step examinations (Table 3).

The unadjusted predictive validity coefficients of the MCAT for USMLE Step 1 through Step 3 examinations range from $r = 0.38$ to 0.60. The restriction of range adjustments increased the correlation coefficients for the biological sciences, physical sciences, and verbal reasoning subtests (e.g., from $r = 0.27$ to 0.34 for USLME Step 1, and from $r = 0.27$ to 0.34 for USLME Step 2) slightly (≤ 0.10). The writing sample subtest correlations with medical board licensing examinations are near zero. The unadjusted effect sizes are illustrated in the random- and fixed-effects model Forrest plots for the MCAT subtests and the USMLE Step 1 examination (Fig. 1). As shown by the different sizes of the plots, the studies are weighted by their respective sample sizes and represented by lines to illustrate their 95% confidence intervals. Although the random-effects size analyses provide a more conservative estimate, the values obtained for each of the subtests are very similar to the fixed-effects analyses. In particular, the random-effects analyses for the biological and physical sciences subtests on the USMLE Step 1 were nearly identical at

Table 2

Random-Effects Model of the MCAT Predictive Validity Coefficients (*r*) With Medical School Performance Measures, from a Meta-Analysis of Studies of the MCAT, 1991–2006

Test	No. of studies (FD)*	Sample size	Basic science/ preclinical†	No. studies (FD)*	Sample size	Clerkship/ clinical†
1991 to 2006	8			4		
MCAT	6 (1)	7,419	0.39 (0.21–0.54) 0.43‡	3 (46)	6,215	0.34 (0.29–0.39) 0.39
Biological sciences subtest	3 (4)	990	0.32 (0.21–0.42) 0.40	1 (1)	275	0.12 (0.00–0.23) 0.15
Physical sciences subtest	3 (1)	990	0.23 (0.09–0.36) 0.26	1 (NA)	275	0.06 (−0.05–0.18) 0.07
Verbal reasoning subtest	3 (5)	990	0.19 (0.12–0.25) 0.24	1 (1)	275	0.14 (0.02–0.25) 0.18
Writing sample subtest§	1 (NA)	126	−0.13 (−0.30–0.05)	1 (NA)	275	0.07 (−0.05–0.19)

* Rosenthal's file drawer method is used to determine the number of unpublished studies (FD) with a mean observed effect of zero that would be needed to make the effect size no longer significant. NA = not applicable.

† Mean (95% confidence interval), SD.

‡ Adjusted effect size for restriction of range[32] $= \rho_i^2 = \dfrac{\rho_i^2(\sigma_1/\sigma_1)^2}{1 + \rho_i^2(\sigma_1/\sigma_1)^2 - \rho_i^2}$

§ The writing sample subtest uses a letter system where standard deviations are not quantified numerically by the Association of American Medical Colleges. Standard deviations for the other subtests were acquired from (www.aamc.org/data/facts/2004/mcatgpabymaj1.htm).

Table 3

Random-Effects Model of the MCAT Predictive Validity Coefficients (*r*) With Medical Board Licensing Examination Measures, from a Meta-Analysis of Studies of the MCAT, 1991–2006

Test	No. of studies (FD)*	Sample size	USMLE Step 1†	No. of studies (FD)*	Sample size	USMLE Step 2†	No. of studies (FD)*	Sample size	USMLE Step 3†
1991 to 2006	16			10			3		
MCAT	6 (33)	29,701	.60 (0.50–0.67) 0.66‡	4 (9)	27,044	.38 (0.26–0.49) 0.43	2 (15)	25,214	.43 (0.32–0.54) 0.48
Biological sciences subtest	9 (44)	15,508	.48 (0.41–0.54) 0.58	5 (1)	3,044	.30 (0.20–0.39) 0.38	1 (2)	650	.11 (0.03–0.19) 0.14
Physical sciences subtest	8 (138)	13,568	.47 (0.43–0.51) 0.52	4 (1)	1,384	.25 (0.03–0.46) 0.28	—	—	—
Verbal reasoning subtest	9 (7)	15,508	.27 (0.19–0.35) 0.34	6 (28)	3,096	.27 (0.22–0.32) 0.34	2 (16)	694	.27 (0.20–0.34) 0.34
Writing sample subtest	7 (1)	13,732	.08 (0.02–0.14)	4 (NA)	2,216	.05 (−0.02 to 0.12)	—	—	—

* Rosenthal's file drawer method is used to determine the number of studies in the fugitive literature (FD) with a mean observed effect of zero that would be needed to make the effect size no longer significant. NA = not applicable.

† Mean (95% confidence interval), SD.

‡ Adjusted effect size for restriction of range[32] $= \rho_1^2 = \dfrac{\sigma_1^2(\sigma_1/\sigma_1)^2}{1 + \rho_1^2(\sigma_1/\sigma_1)^2 - \rho_1^2}$

§ The writing sample subtest uses a letter system where standard deviations are not quantified numerically by the Association of American Medical Colleges. Standard deviations for the other subtests were acquired from (www.aamc.org/data/facts/2004/mcatgpabymaj1.htm).

$r = 0.48$ (95% CI, 0.41–0.54) and $r = 0.47$ (95% CI, 0.43–0.51), respectively, and at $r = 0.27$ (95% CI, 0.19–0.35) for the verbal reasoning subtest. The writing sample showed low predictive validity at $r = 0.08$ (95% CI, 0.02–0.14). As shown in Table 3, restriction of range adjustments were calculated for the MCAT subtests (except the writing sample) across the USMLE Step 1 to Step 3 examinations. For example, on the USMLE Step 1 examination, biological sciences ($r = 0.48$), physical sciences ($r = 0.47$), and verbal reasoning ($r = 0.27$) subtests showed a slight increase in coefficients to $r = 0.58$, $r = 0.52$, and $r = 0.34$, respectively. Although the Cochran *Q* test shows significant heterogeneity between studies on the USMLE Step 1, an analysis to determine the potential differences as a result of moderator variables (e.g., sex, race, ethnicity) was limited by the information collected and reported in each of the studies we included in the meta-analysis. A separate URM subgroup analysis, however, showed an almost identical random-effects size for three studies ($n = 443$) on the MCAT total with the USMLE Step 1, $r = 0.59$ (95% CI, 0.53–0.65).

Discussion

The major findings of the present study are as follows:

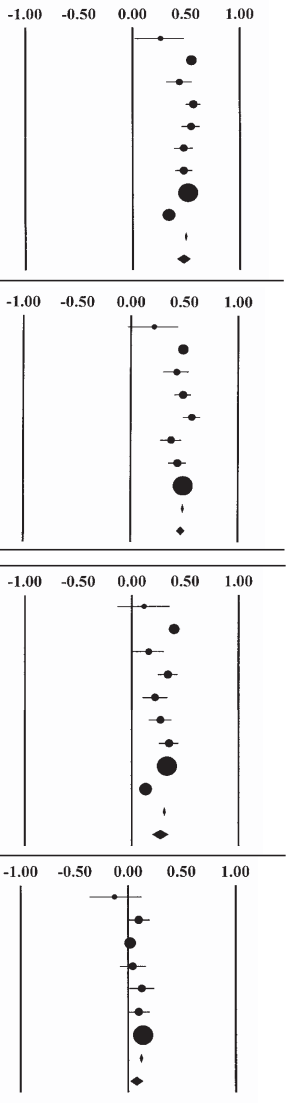
- The MCAT total has an adjusted medium predictive validity coefficient effect size for basic science/preclinical ($r = 0.43$; 18.5% of the variance) and clerkship/clinical ($r = 0.39$; 15.2% of the variance).
- Other than the biological sciences subtest with basic science/preclinical ($r = 0.40$), all of the MCAT subtest predictive validity coefficients are small (range: $r = 0.00$ to 0.29) for both the basic science/preclinical and clerkship/clinical years of medical school performance.
- The MCAT total has a large predictive validity coefficient ($r = 0.66$; 43.6% of the variance) effect size for USMLE Step 1, and medium validity coefficients for USMLE Step 2 ($r = 0.43$; 18.5% of the variance) and USMLE Step 3 ($r = 0.48$; 23.0% of the variance).
- The unadjusted (writing sample) and adjusted MCAT subtest predictive validity coefficients are small to large (range: $r = 0.05$ to 0.58) for the various medical board licensing examinations.
- The writing sample has little predictive validity for both the medical board licensing exam and medical school performance measures.

The MCAT—as a total score and subtests—has small to medium predictive validity of performance in medical school during both the preclinical and clinical years. It is somewhat better at predicting performance on medical licensing exams, accounting for 44% of the variance overall on the USMLE Step 1 and 19% on the USMLE Step 2. The discrepancy in the predictive validity of the MCAT and its subtests between medical school and licensing exam performance is probably attributable to domain and method specificity. Both the MCAT and licensing exams assess knowledge and cognition (the domain) by employing objective testing methods (usually multiple-choice questions), whereas during medical school, other domains (e.g., practical skills) using various methods (e.g., observations, checklists) are assessed. The predictive validity of the MCAT is similar to comparable tests. For the Law School Admission Test (LSAT), for example, *r* ranges from 0.09 to 0.58 (median = 0.41) with law school grades,[15] and for the Graduate Record Examinations (GRE) scores for graduate students, $r = 0.34$ with grade point averages.[16]

There are several caveats to consider in the present study. As for any meta-analysis, our findings are only as good as the quality of the original studies that we

Random and Fixed Effects Model Forrest Plots of the Correlations Between the Medical College Admissions Test (MCAT) Subtests and the United States Medical Licensing Examination (USMLE) Step 1

Study source	Sample size	Weighted mean difference (95% CI)
Biological sciences		
Baker et al,[2] 2000	63	.26 (.02 - .48)
Basco et al,[23] 2002	933	.55 (.51 – .60)
Dixon,[24] 2004	174	.44 (.31 - .55)
Gilbert et al,[26] 2002	355	.57 (.50 - .64)
Kasuya et al,[30] 2003	258	.55 (.46 - .63)
Peterson et al,[35] 2003	285	.48 (.39 - .57)
Simon et al,[36] 2002	355	.48 (.40 - .56)
Swanson et al,[4] 1996	11,145	.52 (.51 - .53)
Veloski,[38] 2000	1,940	.34 (.30 - .38)
Fixed	**15,508**	**.50 (.49 - .51)**
Random	**15,508**	**.48 (.41 - .54)**
Physical sciences		
Baker et al,[2] 2000	63	.22 (-.03 - .44)
Basco et al,[23] 2002	933	.49 (.44 – .54)
Dixon,[24] 2004	174	.43 (.30 - .54)
Gilbert et al,[26] 2002	355	.49 (.41 - .57)
Kasuya et al,[30] 2003	258	.57 (.49 - .65)
Peterson et al,[35] 2003	285	.38 (.28 - .48)
Simon et al,[36] 2002	355	.44 (.35 - .52)
Swanson et al,[4] 1996	11,145	.49 (.47 - .50)
Fixed	**13,568**	**.49 (.47 - .50)**
Random	**13,568**	**.47 (.43 - .51)**
Verbal reasoning		
Baker et al,[2] 2000	63	.11 (-.14 - .35)
Basco et al,[23] 2002	933	.40 (.34 – .45)
Dixon,[24] 2004	174	.16 (.01 - .30)
Gilbert et al,[26] 2002	355	.34 (.25 - .43)
Kasuya et al,[30] 2003	258	.22 (.10 - .33)
Peterson et al,[35] 2003	285	.27 (.16 - .37)
Simon et al,[36] 2002	355	.35 (.26 - .44)
Swanson et al,[4] 1996	11,145	.33 (.31 - .35)
Veloski,[38] 2000	1,940	.13 (.09 - .17)
Fixed	**15,508**	**.31 (.29 - .32)**
Random	**15,508**	**.27 (.19 - .35)**
Writing sample		
Baker et al,[2] 2000	63	-.13 (-.36 - .13)
Gilbert et al,[26] 2002	355	.10 (-.00 - .20)
Hojat et al,[3] 2000	1,271	.02 (-.04 - .07)
Kasuya et al,[30] 2003	258	.05 (-.08 - .17)
Peterson et al,[35] 2003	285	.13 (.01 - .24)
Simon et al,[36] 2002	355	.10 (-.00 - .20)
Swanson et al,[4] 1996	11,145	.14 (.12 - .16)
Fixed	**13,732**	**.12 (.11 - .14)**
Random	**13,732**	**.08 (.02 - .14)**



The Cochran *Q*-test for heterogeneity shows significant overall heterogeneity between studies. (Biological sciences, $Q = 97.602$, 8 *df*, $p < .001$; physical sciences, $Q = 16.853$, 7 *df*, $p < .019$; verbal reasoning, $Q = 95.075$, 8 *df*, $p < .001$; writing sample, $Q = 36.415$, 6 *df*, $p < .001$).

Figure 1 Random- and fixed-effects model Forrest plots of the correlations between the MCAT subtests and the USMLE Step 1.

selected. Some of the studies had samples as small as 25 students, whereas others had very large samples as large as 22,495. We tried to account for these imbalances by using a random-effects weighted effect-size analysis and the Cochran *Q* test for heterogeneity between studies. In an attempt to exercise quality control, we included only studies that had been published in refereed journals. All of the included articles, therefore, had been subjected to the standard peer review process. Unfortunately, most of the original studies in our meta-analysis did not analyze or report potentially important moderator variables such as ethnicity, sex, socioeconomic status, or age for us to code and subsequently conduct a systematic moderator variable analysis.

Although the MCAT as a whole shows relatively consistent and good predictive validity findings for performance in both medical school and on licensing examinations, there was considerable variation on the four different subtests. In particular, the biological sciences subtest had the only adjusted medium effect-size value on measures of medical school performance. In predicting performance on the medical board licensing examination measures, only the biological sciences and verbal reasoning subtests maintained adjusted medium effect-size values across the first two and all three Step examination respectively. Just as some of the subtests have shown consistently good predictive validity performance, the writing sample subtest consistently has shown no predictive validity value across medical school or licensure examination performance. The practical implications of these findings for medical schools support the continuing use of the MCAT total score as a predictor of student performance in medical school and beyond. Consideration should be given, however, to weighting or limiting the use of only the biological sciences and verbal reasoning subtests as the two best measures for predicting future medical student success.

Exploring the influence of moderators in the predictive validity of the MCAT was limited by our ability to extract this information from the studies we included in the present study. Although the MCAT provides predictive validity of students' performance in medical school, admission decisions are also influenced by other cognitive measures such as undergraduate GPA. In addition, there have been recent calls[1,17–19] to explore factors other than the MCAT—particularly noncognitive ones that are associated with effective physicians—as potential criteria for selection into medical school. These may include key personal characteristics such as altruism, empathy, integrity, and compassion.[1,20] Important medical professional organizations such as the Accreditation Council for Graduate Medical Education, the American Board of Medical Specialties, and the Royal College of Physicians and Surgeons of Canada have emphasized the multiplicity of physician roles such as medical expert, collaborator, manager, health advocate, scholar, professional, and communicator.[21–22] A challenge to the medical profession, then, is to develop screening and selection methods and devices that supplement the MCAT by focusing on key personal characteristics and the complex nature of physician roles. Nonetheless, the MCAT continues to be a useful assessment tool

in that it has evidence of predictive validity, although it should not be the only criterion used for selection into medical school.

References

1 Albanese MA, Snow MH, Skochelak SE, Huggett KN, Farrell PM. Assessing personal qualities in medical school admissions. Acad Med. 2003;78:313–321.

2 Baker HH, Cope MK, Fisk R, Gorby JN, Foster RW. Relationship of preadmission variables and first- and second-year course performance to performance on the National Board of Osteopathic Medical Examiners' COMLEX-USA Level I examination. JAOA. 2000;100:153–161.

3 Hojat M, Erdmann JB, Veloski JJ, et al. A validity study of the writing sample section of the Medical College Admission Test. Acad Med. 2000;75(10 suppl):S25–S27.

4 Swanson DB, Case SM, Koenig JA, Killian CD. Preliminary study of the accuracies of the old and new Medical College Admission Tests for predicting performance on USMLE Step 1. Acad Med. 1996;71(10 suppl):S25–S30.

5 Julian ER. Validity of the Medical College Admission Test for predicting medical school performance. Acad Med. 2005;80:910–917.

6 Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. JAMA. 2000;238:2008–2012.

7 Julian E, Lockwood J. Predictive validity of the Medical College Admission Test. Contemp Issues Med Educ. 2000;3:1–2.

8 Wolf FM. Meta-Analysis: Quantitative Methods for Research Synthesis. Beverly Hills, Calif: Sage; 1986.

9 DerSimonian R, Laird N. Meta-analysis in clinical trails. Control Clin Trials. 1986;7: 177–188.

10 Cochran WG. The combination of estimates from different experiments. Biometrics. 1954; 10:101–129.

11 Cohen J. Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Erlbaum; 1988.

12 Rosenthal R. The "file drawer problem" and tolerance for null results. Psych Bulletin. 1979;86:638–641.

13 Edelin KC, Ugbolue A. Evaluation of an early medical school selection program for underrepresented minority student. Acad Med. 2001;76:1056–1059.

14 Glass GV, Hopkins KD. Statistical Methods in Education and Psychology. 3rd ed. Boston: Allyn and Bacon; 1996.

15 Wightman LF. Predictive Validity of the LSAT: A National Summary of the 1990–1992 Correlation Studies. Newton, Pa: Law School Admission Council, LSAC Research Reports Services; 1993.

16 Kuncel NR, Hezlett SA, Ones DS. A comprehensive meta-analysis of the predictive validity of the graduate record examinations: implications for graduate student selection and performance. Psych Bulletin. 2001;127:162–181.

17 Ferguson E, James D, Madeley L. Factors associated with success in medical school: systematic review of the literature. BMJ. 2002; 324:952–957.

18 Parlow J, Rothman AI. Personality traits of first year medical students: trends over a six year period. Br J Med Educ. 1974;54:759–765.

19 Cohen J. Facing the future. President's address presented at: 112th Annual Meeting of the Association of American Medical Colleges; November 4, 2001; Washington, DC.

20 The Medical School Objectives Writing Group. Learning objectives for medical student education—guidelines for medical schools: report I of the Medical School Objectives Project. Acad Med. 1999;74:13–18.

21 Societal Needs Working Group. CanMEDS 2000 project. Skills for the new millennium. Ann R Coll Physicians Surg Can. 1996;29: 206–216.

22 Accreditation Council for Graduate Medical Education. Graduate Medical Education Directory 2001–2002. Chicago, Ill: AMA; 2001.

23 Basco WT Jr, Way DP, Gilbert GE, Hudson A. Undergraduate institutional MCAT scores as predictors of USMLE Step 1 performance. Acad Med. 2002;77(10 suppl):S13–S16.

24 Dixon D. Relationship between variables of preadmission, medical school performance, and COMLEX-USA Level 1 and 2 performance. JAOA. 2004;100:153–161.

25 Evans P, Goodson LB, Schoffman SI. Relationship between academic achievement and student performance on the comprehensive Osteopathic Medical Licensing Examination—USA Level 2. JAOA. 2003;103:331–336.

26 Gilbert GE, Basco WT Jr, Blue AV, O'Sullivan PS. Predictive validity of the Medical College Admissions Test writing sample for the United States Medical Licensing Examination Step 1 and 2. Adv Health Sc Educ. 2002;7: 191–200.

27 Giordani B, Edwards AS, Segal SS, Gillum LH, Lindsay A, Johnson N. Effectiveness of a formal post-baccalaureate pre-medicine program for underrepresented minority students. Acad Med. 2001;76:844–848.

28 Haist SA, Witzke DB, Quinlivan S, Murphy-Spencer A, Wilson JF. Clinical skills as demonstrated by a comprehensive clinical performance examination: who performs better—men or women? Adv Health Sci Educ. 2003;8:189–199.

29 Huff KL, Koenig JA, Treptau MM, Sireci SG. Validity of MCAT scores for predicting clerkship performance of medical students grouped by sex and ethnicity. Acad Med. 1999;74(10 suppl):S41–S44.

30 Kasuya RT, Naguwa GS, Guerrero PS, Hishinuma ES, Lindberg MA, Judd NK. USMLE performances in a predominantly Asian and Pacific Islander population of medical students in a problem-based learning curriculum. Acad Med. 2003;78:483–490.

31 Kossoff EH, Hubbard TW, Gown CW Jr. Early clinical experience enhances third-year pediatrics clerkship performance. Acad Med. 1999;74:1238–1241.

32 Kulatunga-Moruzi C, Norman GR. Validity of admissions measures in predicting performance outcomes: the contribution of cognitive and non-cognitive dimensions. Teach Learn Med. 2002;14:34–42.

33 Mitchell K, Haynes R, Koenig JA. Assessing the validity of the updated Medical College Admission Test. Acad Med. 1994;69:394–401.

34 Ogunyemi D, Taylor-Harris DS. NBME obstetrics and gynecology clerkship final examination scores. J Reprod Med. 2004;49: 978–982.

35 Peterson CA, Tucker RP. Medical gross anatomy as a predictor of performance on the USMLE Step 1. Anat Rec (Part B: New Anat). 2005;283B:5–8.

36 Simon SR, Volkan K, Hamann C, Duffey C, Fletcher SW. The relationship between second-year medical students' OSCE scores and USMLE Step 1 scores. Med Teach. 2002; 24:535–539.

37 Spellacy WN. The OB/GYN clerkship rotation sequence: does it affect performance on final examinations? J Reprod Med. 1998; 43:141–143.

38 Veloski JJ, Callahan CA, Xu G, Hojat M, Nash DB. (2000). Prediction of students' performances on licensing examinations using age, race, sex, undergraduate GPA, and MCAT scores. Acad Med. 2000;75(10 suppl): S28–S30.

39 Violato C, Donnon T. Does the Medical College Admission Test predict clinical reasoning skills? A longitudinal study employing the Medical Council of Canada clinical reasoning examination. Acad Med. 2005;80(suppl 10):S14–S16.

40 Wiley A, Koenig JA. The validity of the Medical College Admission Test for predicting performance in the first two years of medical school. Acad Med. 1996;71(suppl 10):S83–S85.