MCAT Verbal Reasoning score: less predictive of medical school performance for English language learners

Babbi Winegarden,[1] Dale Glaser,[2] Alan Schwartz[3] & Carolyn Kelly[4]

CONTEXT Medical College Admission Test (MCAT) scores are widely used as part of the decision-making process for selecting candidates for admission to medical school. Applicants who learned English as a second language may be at a disadvantage when taking tests in their non-native language. Preliminary research found significant differences between English language learners (ELLs), applicants who learned English after the age of 11 years, and non-ELL examinees on the Verbal Reasoning (VR) sub-test of the MCAT. The purpose of this study was to determine if relationships between VR sub-test scores and measures of medical school performance differed between ELL and non-ELL students.

METHODS Scores on the MCAT VR sub-test and student performance outcomes (grades, examination scores, and markers of distinction and difficulty) were extracted from University of California San Diego School of Medicine admissions files and the Association of American Medical Colleges database for 924 students who matriculated in 1998–2005 (graduation years 2002–2009). Regression models were fitted to determine whether MCAT VR sub-test scores predicted medical school performance similarly for ELLs and non-ELLs.

RESULTS For several outcomes, including pre-clerkship grades, academic distinction, US Medical Licensing Examination Step 2 Clinical Knowledge scores and two clerkship shelf examinations, ELL status significantly affects the ability of the VR score to predict performance. Higher correlations between VR score and medical school performance emerged for non-ELL students than for ELL students for each of these outcomes.

CONCLUSIONS The MCAT VR score should be used with discretion when assessing ELL applicants for admission to medical school.

Medical Education 2012: ***46****: 878–886*
doi:10.1111/j.1365-2923.2012.04315.x

Discuss ideas arising from this article at www.mededuc.com 'discuss'

[1]Division of Medical Education and Department of Psychiatry, University of California San Diego, La Jolla, California, USA
[2]Adjunct Faculty, University of San Diego, San Diego, California, USA
[3]Department of Medical Education, University of Illinois at Chicago, Chicago, Illinois, USA
[4]Division of Medical Education and Department of Medicine, University of California San Diego, La Jolla, California, USA

Correspondence: Babbi Winegarden, Department of Psychiatry, University of California San Diego (UCSD), 9500 Gilman Drive, 301 University Center, Room 102, La Jolla, California 92093-0092, USA. Tel: 00 1 858 534 0362; Fax: 00 1 858 534 0595; E-mail: bwinegarden@ucsd.edu

INTRODUCTION

Gathering validity evidence for tests such as the Medical College Admission Test (MCAT) is critical for determining whether such test scores may be used to predict future performance in medical and health professions schools. Assessing whether predictive relationships are consistent across subgroups is an important component of this validation process. The *Standards for Educational and Psychological Testing* advise test developers and users of test results to be aware of possible score differences between 'relevant subgroups', particularly when the 'life chances or educational opportunities' of an examinee in one of those subgroups may be significantly affected.[1] If test data show significant differences between subgroups, the *Standards* require an investigation into such differences to determine whether the test is really measuring what it purports to measure for each of the relevant subgroups (Standard 7.10).[1]

How standardised admissions tests are being used as part of admissions processes is under scrutiny in many countries around the world, including the USA. Predictive validity concerns are being raised with regard to the Undergraduate Medicine and Health Sciences Admission Test (UMAT),[2] the Health Professions Admission Test (HPAT)[3] and the Graduate Medical School Admissions Test (GAMSAT).[4] As the Association of American Medical Colleges (AAMC) is planning to release a new version of the MCAT in 2015, it is important to examine issues related to its predictive validity.

The AAMC MCAT is a widely used and important part of admissions processes for medical and health professions schools in the USA, Canada and many other countries. Validity evidence for its use in predicting medical school performance has been demonstrated. Donnon *et al.*[5] performed a meta-analysis of 23 published articles on the ability of the MCAT to predict performance in medical school and on US Medical Licensing Examination (USMLE) components. They concluded that MCAT scores have a predictive validity that ranges from small to medium for various measures of performance in medical school and on the USMLE components. Callahan *et al.*[6] looked at the predictive validity of the MCAT across its last three versions and concluded that there was broad support for its predictive validity in general for performance in medical school and on USMLE components.

The relationship between MCAT scores and academic distinction and difficulty in medical school appears to be more variable.[7] There is a general trend towards an increased likelihood of distinction with higher MCAT scores, as well as an increased likelihood of academic difficulty with lower MCAT scores. However, these relationships are not perfect. Julian[7] notes that student characteristics that are unrelated to knowledge, skills or abilities may affect these outcomes. For example, 11% of students with MCAT Verbal Reasoning (VR) sub-test scores of < 4 experienced academic difficulty; however, the other 89% of students with VR scores of < 4 did not experience academic difficulty.

For any test, facility with the test's language may affect performance.[8] Abedi and Gándara[9] assert that the assessment of English language learners (ELLs) can be difficult. They propose that the complexity of the language used for assessment can significantly hamper ELLs because they differ not only in their language background, but also in their 'cultural, family and personal characteristics'.[9] To date, no-one has investigated the impact of learning English at a later age on the predictive validity of the MCAT VR sub-test score for performance in medical school or for USMLE scores.

The purpose of the current study was to determine whether MCAT VR sub-test scores predict performance in medical school and on USMLE components similarly for ELL and non-ELL students. Based on publicly available data for the MCAT, in which learning English as a second language was defined as 'learning English after the age of 11 [years]', and after simple analyses, we found that ELLs perform more poorly as a group on the MCAT VR sub-test. Given these group differences, it is important to determine if the MCAT VR score predicts performance in medical school similarly for ELLs and non-ELLs. Additionally, we hypothesised that the learning of English as a second language may be responsible for this difference in performance and may represent construct-irrelevant variance; in other words, performance on the MCAT VR sub-test for applicants who learned English after the age of 11 years may be affected by the fact that they are taking a test in a non-native language and may not be truly representative of the construct of verbal reasoning *per se*. If this is true, then VR scores for ELLs may not be as predictive of performance in medical school as they are for non-ELL candidates. If the VR score is not as predictive for ELLs as it is for non-ELLs, care must be applied to the use of the MCAT VR score in admissions processes for ELLs.

METHODS

Participant data

Data were gathered retrospectively from the University of California San Diego (UCSD) School of Medicine and AAMC archives. Data were de-identified for analysis, encrypted as the institutions exchanged data, and kept in a secure location. An examinee's ELL status was defined according to the age at which he or she had first learned English. This variable was gathered from the AAMC MCAT archives. The MCAT archives used a cut-off descriptor of 'learned English after the age of 11' and thus we divided examinees into two groups. The ELL group included examinees who had learned English after the age of 11 years and the non-ELL group included examinees who had learned English before the age of 11 years. The other predictor variable was the score on the MCAT VR sub-test. This variable and the outcome variables were gathered from the UCSD School of Medicine archives and included data for students in matriculation years 1998–2005, graduation years 2002–2009. Inter-institution agreements for data sharing were obtained. The study was approved by the UCSD Human Research Protections Program (project: 090175) and the University of Illinois at Chicago Office for the Protection of Research Subjects (protocol: 2009-0199).

Outcome variables

The outcome variables included:

1. weighted average grade score (Fail-F [weight = 1], Remediate-Y [weight = 2], Pass-P [weight = 3], Honours-H [weight = 4]) in the pre-clerkship or first 2 years of medical school courses;
2. weighted average grade score (F, Y, P, H) for clerkships or Year 3 of medical school;
3. academic difficulty and academic distinction, where academic difficulty is represented by the number (count) of Ys or Fs obtained by the student in the first 3 years of medical school (students who receive Ys need to remediate some portion of the course; students who receive Fs need to repeat the course) and academic distinction is represented by the number (count) of Hs obtained in the first 3 years of medical school;
4. clerkship shelf examination scores on first examinations in medicine, neurology, paediatrics, psychiatry, reproductive medicine and surgery;
5. USMLE Step 1 score on first examination attempt;
6. USMLE Step 2 Clinical Knowledge (CK) score on first examination attempt, and
7. USMLE Step 2 Clinical Skills (CS) pass/fail on first examination attempt.

Data analysis

A hierarchical approach was used for testing a moderated multiple regression model. For each outcome variable, we first entered ELL status (ELL or non-ELL) and MCAT VR scores as predictors, and then, in a second step, entered the interaction between the two predictors into the model. Scores on the VR sub-test were centred in order to minimise non-essential collinearity.[10,11]

Linear regressions were performed when outcomes were continuous and expected to be normally distributed. When the outcome variable was binary (e.g. pass/fail), multiple logistic regression was conducted. These analyses were conducted using SPSS Version 18.0.2 (SPSS, Inc., Chicago, IL, USA). When outcomes were counts of unusual events (e.g. number of failed courses), zero-inflated Poisson regression was performed using Mplus Version 6.1 (Muthén & Muthén, Los Angeles, CA, USA).

In these models, the key test for the differential predictive ability of the VR sub-test on medical school outcomes for ELLs and non-ELLs is whether the interaction term is significant. In the coding of our data, a significant negative interaction would be interpreted as a finding that the relationship between VR score and outcome is stronger for non-ELLs than for ELLs. When interactions were significant, we examined the associations between VR score and outcome within each group.

RESULTS

Participants

The original dataset consisted of 971 students enrolled at UCSD School of Medicine from 1998 through to 2005. A total of 47 students were excluded because the critical ELL/non-ELL information was unavailable for them, leaving 924 students in the final dataset. Mean VR score was 8.51 (standard deviation [SD] = 1.70) for the ELL group (n = 72, 7.8%) and 10.13 (SD = 1.56) for the non-ELL group (n = 852, 92.2%). An independent samples t-test resulted in a significant difference ($t_{922} = 8.38$, $p < 0.001$).

Pre-clerkship (Years 1 and 2) and clerkship (Year 3) weighted average

With weighted average pre-clerkship grades as the outcome variable, at the first step VR score and ELL status accounted for 3.4% of the variance ($F_{2,913}$ = 15.84, p < 0.001). Adding the interaction term improved the model fit by 1.3% ($F_{1,912}$ = 12.33, p < 0.05), although the increment to the model fit is small: r^2 change = 0.013 (1.3% incremental variance). Score on the VR sub-test was significant at the first step (b = 0.018, p < 0.001). All predictors were significant in the final model (i.e. ELL: b = − 0.089, p = 0.001; VR: b = 0.022, p < 0.001), including the ELL × VR interaction term (b = − 0.044, p < 0.001). Table 1 summarises these results. Note that there are some large standardised residuals (although there were no particularly high values for Cook's distance), and further model fit may be a future consideration.

Figure 1 shows the regression lines for the non-ELL and ELL groups for the pre-clerkship weighted average and helps shed insight into the nature of the interactions. For pre-clerkship weighted average, a negative slope is found for the ELL group and a positive slope for the non-ELL group. The significant interaction suggests that the VR score yields a differential pattern of predicting performance in the first 2 years of medical school in the two groups.

With weighted average clerkship grades as the outcome variable, at the first step VR score and ELL status accounted for 7.6% of the variance ($F_{2,867}$ = 35.78, p < 0.001), representing a significant addition to model fit. At the second step the interaction term did not significantly add to model fit ($F_{1,866}$ = 3.38, p = 0.066). The increment to the model fit is very small: r^2 change = 0.004 (0.4% incremental variance, meaning that < 0.5 of 1% of the incremental variance is attributable to the addition of the multiplicative term).

When examining the individual predictors at the first step, only the centred score for VR is significant: b = 0.039 (p < 0.05). At the second step, both predictors but not the interaction term are significant: b = − 0.033 (p = 0.066). The zero-order correlation of the interaction term with the criterion is r = 0.09 and the partial correlation = − 0.062 (Table 1). Note that there are some large standardised residuals (although there were no particularly high values for Cook's distance); hence, future analyses may entail either transformation of the criterion variable or reanalysis without the largest standardised residuals.

Academic distinction (total H) and difficulty (total Y and F)

The academic distinction (total number of Honours grades) and difficulty (total number of Remediate or Fail grades) outcomes were examined using zero-inflated Poisson regression for the Poisson model. Information theoretic indices such as the Akaike

Table 1 Linear regression results for pre-clerkship and clerkship grade point average (GPA)

	Pre-clerkship GPA				Clerkship GPA			
		Unstandardised coefficients		Standardised coefficients		Unstandardised coefficients		Standardised coefficients
Predictor	ΔR^2	B	SE	β	ΔR^2	B	SE	β
Step 1								
ELL	0.034	− 0.029	0.022	− 0.044	0.076	− 0.043	0.031	− 0.047
VR		0.018‡	0.004	0.166		0.039‡	0.005	0.26
Step 2								
ELL	0.013	− 0.089‡	0.028	− 0.136	0.004	− 0.088*	0.04	− 0.096
VR		0.022‡	0.004	0.203		0.042‡	0.005	0.279
ELL × VR		− 0.044‡	0.013	− 0.157		− 0.033	0.018	− 0.083

* p < 0.05; † p < 0.01; ‡ p < 0.001
SE = standard error; ELL = English language learner; VR = Verbal Reasoning sub-test

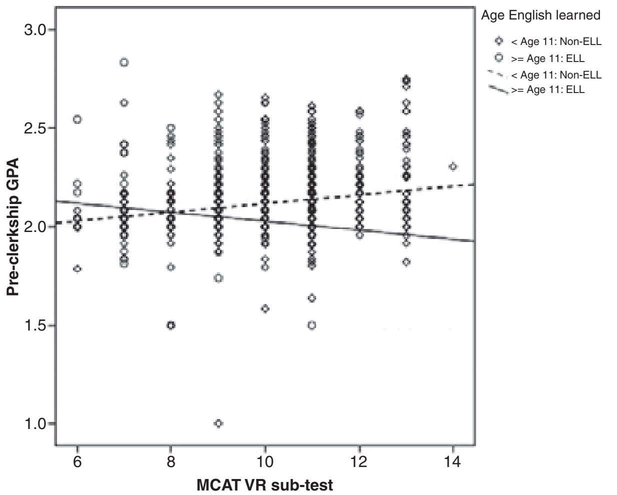

Figure 1 Regression lines for language status for Medical College Admission Test Verbal Reasoning sub-test (MCAT VR) score and pre-clerkship grade point average (GPA). ELL = English language learner

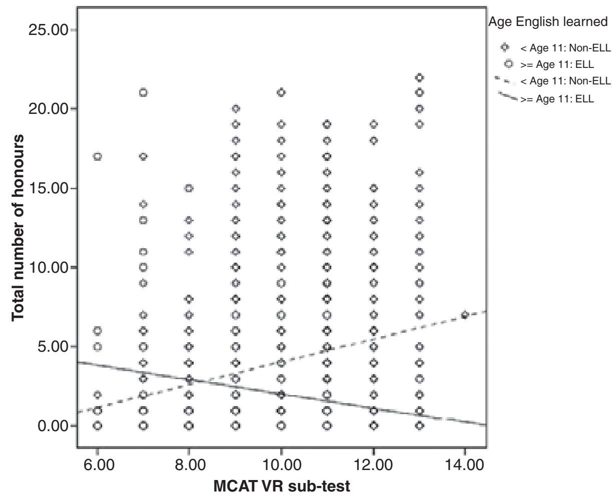

Figure 2 Zero-inflated Poisson regression for total number of honours (Years 1–3). MCAT VR = Medical College Admission Test Verbal Reasoning sub-test; ELL = English language learner

information criterion (AIC) and Bayesian information criterion (BIC) were used to compare models with and without the interaction term (smaller AIC and BIC values indicate improved fit given the number of parameters included).

For academic distinction, in the initial model without the interaction term, log-likelihood = − 2865.49, AIC = 5743 and BIC = 5771. In this model, VR score was significant for the Poisson ($b = 0.106$) and logistic ($b = -0.198$) components, but ELL status was not a significant predictor in either portion of the model.

For the full model including interaction, log-likelihood = − 2840.99, AIC = 5698 and BIC = 5737, suggesting improved fit. All of the predictors were significant for the Poisson component, including the interaction term ($b = -0.313$, $p = 0.001$). The only significant term for the logistic portion of the model was VR ($b = -0.218$, $p < 0.001$).

Figure 2 illustrates the model findings and shows a positive relationship between VR score and predicted total Honours in the non-ELL group, and a negative relationship between VR score and predicted total Honours in the ELL group.

For the outcome of academic difficulty, neither model had any significant predictors.

Clerkship shelf examinations

An overall positive main effect of VR emerged in all shelf examinations. However, as Table 2 shows, the interaction between VR score and ELL status was significant and negative for shelf examinations in paediatrics and psychiatry. On the paediatrics examination, the inclusion of the interaction term ($b = -1.63$, $p = 0.032$) improved the model slightly but significantly ($\Delta R^2 = 0.007$, $F_{1,558} = 4.64$, $p = 0.032$). In psychiatry, inclusion of the interaction term ($b = -2.35$, $p = 0.003$) also improved the model ($\Delta R^2 = 0.014$, $F_{1,560} = 8.83$, $p = 0.003$). The interaction was not significant for shelf examinations in medicine ($b = -0.98$, $p = 0.183$), neurology ($b = -0.748$, $p = 0.291$), reproductive medicine ($b = -0.88$, $p = 0.225$) or surgery ($b = -1.01$, $p = 0.219$).

As Fig. 3 shows, on the paediatrics shelf examination, a positive association emerged between VR and examination scores for non-ELL students, but not for ELL students. On the psychiatry shelf examination, there was a positive association between VR and examination scores for non-ELL students and a negative association for ELL students.

USMLE Step 1 and 2 scores

USMLE Step 1 and Step 2 CK (total scores)

With USMLE Step 1 total score as the criterion variable, at the first step (individual predictors only) 8.5% of the variance was accounted for ($F_{2,906} = 41.39$, $p < 0.001$). Adding the interaction term did not significantly add to model fit ($F_{1,905} = 2.75$, $p = 0.098$), and the interaction term was not significant ($b = -2.40$, $p = 0.098$). The zero-order correlation of the interaction term with the

Table 2 Linear regression results for clerkship shelf examinations

	Medicine shelf				Neurology shelf				Paediatrics shelf			
		Unstandardised coefficients		Standardised coefficients		Unstandardised coefficients		Standardised coefficients		Unstandardised coefficients		Standardised coefficients
Predictor	ΔR^2	B	SE	*β*	ΔR^2	B	SE	*β*	ΔR^2	B	SE	*β*
Step 1												
ELL	0.141	– 1.310	1.25	– 0.043	0.115	– 3.581†	1.201	– 0.124	0.115	– 2.868*	1.309	– 0.091
VR		1.769‡	0.201	0.360		1.294‡	0.193	0.280		1.544‡	0.213	0.301
Step 2												
ELL	0.003	– 2.874	1.713	– 0.094	0.002	– 4.774†	1.649	– 0.166	0.007	– 5.320†	1.731	– 0.169
VR		1.849‡	0.21	0.377		1.354‡	0.201	0.293		1.685‡	0.223	0.329
ELL × VR		– 0.982	0.736	– 0.078		– 0.748	0.708	– 0.063		– 1.629*	0.756	– 0.124
	Psychiatry shelf				Reproductive medicine shelf				Surgery shelf			
Step 1												
ELL	0.129	– 3.858†	1.346	– 0.119	0.12	– 1.332	1.225	– 0.045	0.090	– 1.267	1.420	– 0.037
VR		1.607‡	0.219	0.304		1.562‡	0.197	0.330		1.554‡	0.225	0.288
Step 2												
ELL	0.014	– 7.620‡	1.840	– 0.235	0.002	– 2.713	1.671	– 0.092	0.002	– 2.755	1.864	– 0.081
VR		1.801‡	0.227	0.341		1.633‡	0.205	0.345		1.636‡	0.234	0.303
ELL × VR		– 2.353‡	0.791	– 0.176		– 0.876	0.722	– 0.072		– 1.012	0.822	– 0.070

* $p < 0.05$; † $p < 0.01$; ‡ $p < 0.001$
SE = standard error; ELL = English language learner; VR = Verbal Reasoning sub-test

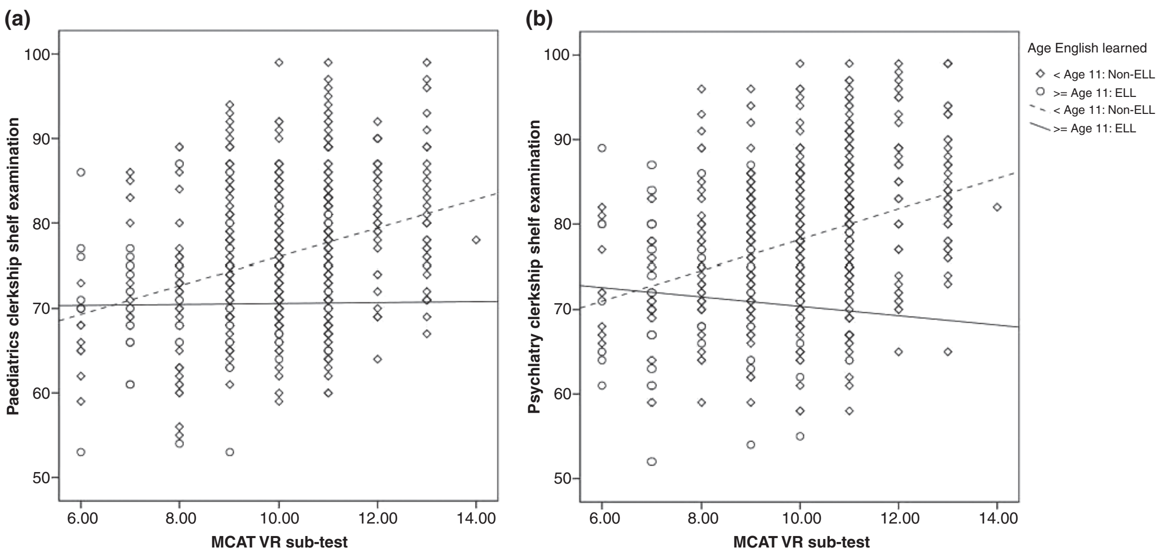

Figure 3 Regression lines for language status for scores on the Medical College Admission Test Verbal Reasoning sub-test (MCAT VR) and the (a) paediatrics and (b) psychiatry clerkship shelf examinations. ELL = English language learner

criterion is $r = 0.097$ and the partial correlation is -0.055. Again, some large standardised residuals emerged (although there were no particularly high values for Cook's distance).

For USMLE Step 2 CK total score, at the first step (individual predictors only) 8.6% of the variance was accounted for ($F_{2,847} = 40.90$, $p < 0.001$). Adding the interaction term significantly improved model fit ($F_{1,846} = 6.25$, $p = 0.013$); the interaction was significant and negative ($b = -4.25$, $p = 0.013$). Table 3 summarises these results.

Figure 4 shows a positive association between VR and USMLE Step 2 CK scores in the non-ELL group, but no association in the ELL group.

USMLE Step 2 CS (pass/fail on first attempt)

There were no significant predictors of failing the USMLE Step 2 CS component in the model with or without the interaction. The inability to predict this outcome is probably attributable to a very small number of examinee failures in the data ($n = 4$).

Figure 5 summarises the differences in correlations between VR scores and all hypothesised medical school outcomes for ELLs and non-ELLs. In all cases, lower correlations are observed for ELLs than for non-ELLs; in most cases, these correlations are also of smaller absolute magnitude (i.e. weaker).

DISCUSSION

The MCAT VR sub-test score is often used as a predictor of future academic performance by admissions committees when evaluating applicants for medical and health professions schools. When used as part of an algorithm[12] or as part of a holistic process, the VR score can influence whether an applicant is considered for an institution's specific secondary application, an interview or, ultimately, for admission.

Much previous research supports the use of the VR score and many of the other MCAT subscores in predicting future medical school performance.[5,6] In the present study, for students who learned English before the age of 11 years, we generally found significant positive associations between VR score and medical school performance, and thus the study supports this conclusion. However, for students who learned English after the age of 11 years, the predictive ability of the VR score was found to be substantially lower (and sometimes negative). Although the effect sizes for some of the findings were relatively small, there is a body of literature that illustrates how even small effects may yield some import as their impact depends on the nature of the design, what was anticipated and the context.[13,14] Importantly for this study, the impact of the VR score on ultimate admission to medical school is fairly

Table 3 Linear regression results for US Medical Licensing Examination Steps 1 and 2

	Step 1 score				Step 2 Clinical Knowledge score			
		Unstandardised coefficients		Standardised coefficients		Unstandardised coefficients		Standardised coefficients
Predictor	ΔR^2	**B**	**SE**	β	ΔR^2	**B**	**SE**	β
Step 1								
ELL	0.085	– 3.096	2.532	– 0.040	0.088	– 1.648	2.962	– 0.019
VR		3.440‡	0.409	0.277		4.095‡	0.477	0.291
Step 2								
ELL	0.003	– 6.450*	3.239	– 0.084	0.007	– 7.344*	3.729	– 0.084
VR		3.650‡	0.428	0.294		4.459‡	0.498	0.317
ELL × VR		– 2.403	1.449	– 0.073		– 4.250*	1.700	– 0.112

* p < 0.05; † p < 0.01; ‡ p < 0.001
SE = standard error; ELL = English language learner; VR = Verbal Reasoning sub-test

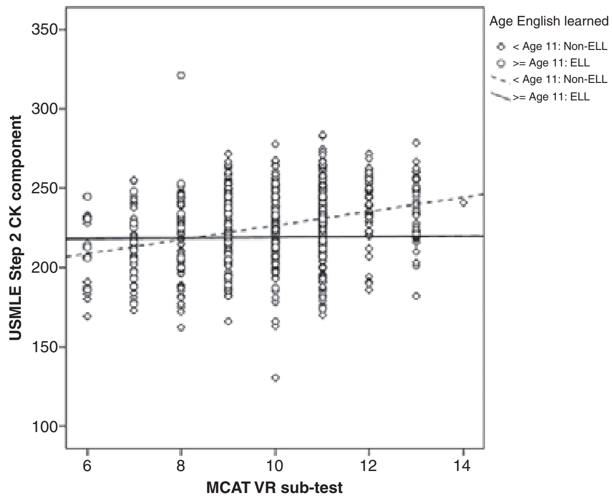

Figure 4 Regression lines for language status for scores on the Medical College Admission Test Verbal Reasoning sub-test (MCAT VR) and the US Medical Licensing Examination (USMLE) Step 2 Clinical Knowledge (CK) component. ELL = English language learner

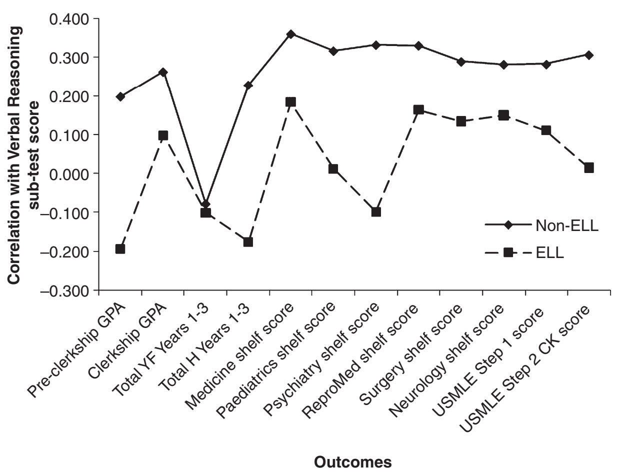

Figure 5 Correlations between verbal reasoning and outcomes based on language status. ELL = English language learner; GPA = grade point average; Y = remediate; F = fail; H = honours; ReproMed = reproductive medicine; USMLE = US Medical Licensing Examination; CK = Clinical Knowledge component

substantive and thus even a small effect size is meaningful.

Our findings show that ELL status significantly affects the ability of the VR score to predict medical school performance in pre-clerkship grades, academic distinction (honours grades), USMLE Step 2 CK scores, and scores on the paediatrics and psychiatry clerkship shelf examinations. Indeed, ELL applicants appear to often outperform this predictor. One possible explanation for this finding in the ELL group may be that admissions committees seek other strengths associated with success in medical school when reviewing ELL applications and thus successful ELL applicants tend to outperform MCAT predictions. It may also be that, once admitted, ELL students may feel the need to overachieve to prove the lower scores incorrect as a measure of prediction. Finally, as a multiple-choice test, the MCAT may not accurately reflect the ability of ELL students to perform in medical school courses designed with clear and structured objectives. These findings deserve further exploration.

The results of this study support the contention that language status with regard to the MCAT VR sub-test score be considered to represent construct-irrelevant variance (the variance in test scores that does not reflect random error or the construct being tested, such as ability to succeed in medical school). When construct-irrelevant variance is shown to exist, care must be taken in the interpretation of those test scores for that group or population. Our finding of the lack of a positive relationship between the MCAT VR score and medical school performance of ELL applicants underscores the importance of taking language status into account when using these scores as part of the admissions process.

At the UCSD medical school, from which the data studied in this research were sourced, approximately 25% of the entering class in 2009 and 2010 had learned English as a second language. Many US medical schools are likely to have even higher percentages of ELL students in their applicant pool than in their matriculating classes. The findings of this study are thus important to both the process of admissions and medical school advisors, who may choose to identify 'at risk' student populations for enhanced academic support.

Future research might focus on a number of areas. Because the literature on age-related language acquisition has changed somewhat over the years, with some continuing support for the 'absoluteness of the age factor' in second language acquisition, as well as 'the notion that there may not be one, but a number, of age-related factors at work',[15] future research might further examine the impact of learning English as a second language and performance on the MCAT VR scale as a function of age by breaking the sample down according to the ages at which the applicants learned English. In addition, this same issue could be explored with reference to performance on the other MCAT subscales. Further, given these findings, other standardised test predictive validity research such as that currently underway for the HPAT, UMAT and GAMSAT may benefit from including language status as part of the study design. Continued work on alternative variables correlated with success in medical school is certainly warranted.

Finally, there are limitations inherent in this study. These findings are based on retrospective data collection from a single institution and therefore may not represent medical school applicants as a whole. In addition, selection bias exists as only students who were actually admitted were studied.

Contributors: BW contributed to the study conception and design, and to the acquisition, analysis and interpretation of data, and drafted the article. DG, AS and CK contributed to the study conception and design, and to the analysis and interpretation of data. All authors contributed to the critical revision of the article and approved the final manuscript for publication.
Acknowledgements: the authors wish to thank the Association of American Medical Colleges for graciously consenting to work on this project and for supplying the data on English language learners.
Funding: this study was supported by the Masters of Health Professions Education programme at the University of Illinois at Chicago, and the University of California San Diego School of Medicine as part of the first author's Master's thesis.
Conflicts of interest: none.
Ethical approval: this study was approved by the University of California San Diego Human Research Protections Program (project: 090175) and the University of Illinois at Chicago Office for the Protection of Research Subjects (protocol: 2009-0199).

REFERENCES

1 American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: AERA 1999.
2 Katelaris AG. What's the matter with the UMAT? *Med J Aust* 2011;**194** (7):330.
3 Halpenny D, Cadoo K, Halpenny M, Burke J, Torreggiani WC. The Health Professions Admission Test (HPAT) score and leaving certificate results can independently predict academic performance in medical school: do we need both tests? *Ir Med J* 2010;**103** (10):300–2.
4 Coates H. Establishing the criterion validity of the Graduate Medical School Admissions Test (GAMSAT). *Med Educ* 2008;**42** (10):999–1006.
5 Donnon T, Paolucci EO, Violato C. The predictive validity of the MCAT for medical school performance and medical licensing examinations: a meta-analysis of the published research. *Acad Med* 2007;**82** (1):100–6.
6 Callahan CA, Hojat M, Veloski J, Erdmann JB, Gonnella JS. The predictive validity of three versions of the MCAT in relation to performance in medical school, residency, and licensing examinations: a longitudinal study of 36 classes of Jefferson Medical College. *Acad Med* 2010;**85** (6):980–7.
7 Julian ER. Validity of the Medical College Admissions Test for predicting medical school performance. *Acad Med* 2005;**80** (10):910–7.
8 Abedi J. Language issues in item development. In: Downing SM, Haladyna TM, eds. *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum Associates 2006;377–398.
9 Abedi J, Gándara P. Performance of English language learners as a subgroup in large-scale assessment: interaction of research and policy. *Educ Meas Issues Pract* 2006;**25** (4):36–46.
10 Cohen J, Cohen P, West SG, Aiken LS. *Applied Multiple Regression/*Correlation Analysis for the Behavioral Sciences, 3rd edn. Mahwah, NJ: Lawrence Erlbaum Associates 2003.
11 Aiken LS, West SG. *Multiple Regression: Testing and Interpreting Interactions*. Newbury Park, CA: Sage Publications 1991.
12 Albanese MA, Snow MH, Skochelak SE, Huggett KN, Farrell PM. Assessing personal qualities in medical school admissions. *Acad Med* 2003;**78** (3):313–321.
13 Prentice D, Miller D. When small effects are impressive. *Psychol Bull* 1992;**112** (1):160–4.
14 Cortina JM, Landis RS. When small effect sizes tell a big story, and when large effect sizes don't. In: Lance CE, Vandenberg RJ, eds. *Statistical and Methodological Myths and Urban Legends*. New York, NY: Routledge 2009; 287–308.
15 Singleton D. Age and second language acquisition. *Annu Rev Appl Linguist* 2001;**21**:77–89.

Received 20 July 2011; editorial comments to authors 1 September 2011, 17 January 2012; accepted for publication 25 April 2012