What Engineering Admissions Can Learn from Medical School Admissions

Harold I. Reiter
Faculty of Health Sciences
McMaster University
Hamilton, Canada
reiterh@mcmaster.ca

Yona Baskharoun
Senior Account Executive
Altus Assessments
Toronto, Canada
ybaskharoun@altusassessments.com

Kelly Dore
Faculty of Health Sciences
McMaster University
Hamilton, Canada
dore@mcmaster.ca

***Abstract*—Medical school admissions have undergone a drastic change in the past decade with increasing awareness about issues of professionalism and a lack of diversity among physicians. These problems are also arising among engineers, where issues of communication, ethics, and cultural competency are beginning to be identified across the profession. Over the years, the traditional admissions process into medical school has been identified to be particularly problematic, as it places a strong emphasis on the cognitive competencies of incoming students, but frequently neglects their non-cognitive competencies or professionalism. The same parallel can be drawn for admission into engineering programs, where the process strongly focuses on cognitive abilities (GPA, standardized test scores) and less so on the non-cognitive skills. In this paper, we suggest various tools to promote the shift to the assessment of professionalism in engineering admissions to ensure that the profession will keep up with the future needs of the population.**

Keywords—admissions; medical education; professionalism; non-cognitive competencies; assessment

I. INTRODUCTION

There have been dramatic changes in medical school admissions over the last 20 years, sharing with engineering schools a parallel direction of burgeoning emphasis placed on professional skills assessment. Medical schools differ from their engineering counterparts in terms of precipitating factors and resource allocation, identification of the challenges in test design, and extent of progress and implementation.

This paper is meant as a guide to the transformation of student selection for medical schools over the last two decades, particularly how the details of that transformation translate to engineering schools. At the outset, Section 1 describes the contrast in precipitating factors and resource allocation, between medical and engineering schools, which in part explains how their common cause towards professional skills assessment has taken different paths. Section 2 will identify the challenges to assessment of professional skills applicable equally to both sets of schools. The challenges are categorized by their source – emanating from test-takers, test parameters, and environmental factors. Section 3 will present assessment tools to face down those challenges, divided between tests that have demonstrated enough data to support their use, and tests that hold at least some promise, but are still too early in development to have garnered enough data to support their use.

II. PRECIPANTS FOR THE ASSESSMENT OF PROFESSIONALISM

While the precipitants for increasing emphasis on student professionalism for both medical and engineering schools is an understandably convoluted affair, the most dominant driver for each is different. For medicine, the chief precipitant has been accountability; for engineering, market forces.

The first critical change towards modern medical schools in the United States was the Carnegie Foundation's 1910 Flexner Report [1]. Among other key recommendations, the Flexner Report launched American medical schools away from homespun remedies towards the objectivity of the scientific method. The scientific method required cognitive skills that homespun remedies did not. Flexner had reported the not uncommon medical school practice of admitting students lacking a high school diploma, whose success on high school "equivalency exams" run by local medical school admissions officers may have had more to do with the applicant's wallet than his/her intellectual capacity. As medical schools changed, so did their admissions criteria, with the implementation of what would later become the Medical College Admissions Test (MCAT) [2].

Once standardized tests of knowledge were instituted in the first half of the 20th century, only high cognitive performers have been accepted to competitive professions like medicine and engineering. The unanticipated corollary became evident

Harold Reiter and Kelly Dore are co-founders and shareholders of *Altus Assessments*, the company responsible for the construction and delivery of the CASPer$^{®}$ test. Yona Baskharoun is an employee of *Altus Assessments*.

by the 1950's – a study in the Annals of Medicine from Washington D.C. reported that the vast majority of complaints against medical practitioners were related to lapses in professionalism [3], a finding punctuated over fifty years later by a more nation-wide study revealing that lapses in professionalism accounted for roughly 95% of medical state boards' disciplinary actions against physicians [4]. During the same era, through political means (prompted by a failed doctors' strike in Ontario in 1986), regional and then national governments influenced national regulatory bodies to define practitioner competencies in a rising wave of documents demanding greater accountability [5].

The first of these was CanMEDS 2000 [6]; its required competencies included medical knowledge but were more predominantly professional competencies – ability to behave ethically, and to act towards patients and co-workers as skilled communicators, collaborators, and managers. CanMEDS 2000 was soon followed by a rising wave of documents focusing on accountability. For medicine in the United States, its parallels came from the Accreditation Committee for Graduate Medical Education (ACGME) [7] and later from the Association of American Medical Colleges (AAMC) [8]. Similar documents have emerged in near lockstep for schools of engineering. In 1996, the ABET engineering criteria, also known as Engineering Criteria 2000 [9], was approved. Of the 11 skills identified, six could more reasonably be described as professional skills (briefly identified as teamwork, ethics, communication, commitment to lifelong learning, and understanding of societal impact and of contemporary issues). More recently, a collaborative endeavor of the United States Department of Labor and the American Association of Engineering Societies describes a multi-tier approach to engineering education [9]. Tier 1 describes the personal effectiveness competencies of integrity, professionalism, dependability and reliability, adaptability and flexibility, interpersonal skills, ability and willingness to learn, and motivation. Tier 2 includes communication, and Tier 3 includes teamwork, customer focus, and problem-solving and decision-making.

The lists of professional skills categorized for both physician and engineer training are remarkably similar. However, the predominant precipitant for engineering schools has not been accountability but rather market forces – supply and demand. The supply of graduating skilled engineers has risen dramatically [10], with significant labor force entry from emerging economies in the Far East (Taiwan, Korea, China) and South Asia (India). At the same time, demand is not so much shrinking as changing. Automation of non-client-facing work previously done by live engineers has shifted greater attention toward client-facing practice; increased project complexity requires greater teamwork between a widening spectrum of co-workers. Both factors have enhanced the value of professionalism.

As Sections 2 (obstacles to assessment of professionalism) and Section 3 (solution for assessment of professionalism) present, medical education studies greatly exceed engineering education studies in defining the overall literature, despite similar goals. There are a number of causes for this discrepancy. While both accountability and market forces are powerful precipitants, the former is more readily treated with a sense of urgency due to sensationalized reports in the media (e.g. Harold Shipman [11]); slow and inexorable market changes rarely merit front page news. Medicine has had the longer lead-in time, commencing in the 1950's [3]. Medical schools commonly have well-resourced divisions of education, often staffed by quantitative, qualitative, and mixed methodology researchers with a diversity of backgrounds in psychology and other humanities. Thankfully, none of the research and development conducted in this area is unique to medical schools. All is applicable to engineering schools, as the upcoming sections make clear.

III. Obstacles to the Assessment of Professionalism

There is a bewildering array of available tests designed to measure professional skills. The options can be greatly simplified by discarding all those that are subject to critical obstacles. To the casual observers' benefit, this will greatly narrow the field of newly emerging measures but may also exclude some traditional tests which have been trusted intuitively. The ten obstacles listed are categorized by their source – test-taker parameters, test parameters, environmental parameters.

A. Test-taker parameters, i.e. Manipulability

1) Response Distortion: When the stakes are high, test-takers will be more prone to distort their responses to present their candidacy in a more favorable light. Surprisingly, not all measures are designed to thwart this activity. At their core, both reference letters and personal statements encourage response distortion. Applicants select referees who are most biased in their favor and seek edits (and sometimes wholesale ghost-writing) to portray themselves with the greatest degree of positive bias in personal statements. Unsurprisingly, therefore, literature reviews have demonstrated an absence of predictive validity for these measures [12]. Less obviously, some test designs simply cannot overcome test-taker ability to distort. When used for low stakes (e.g., career development purposes), Big Five Factor personality testing demonstrates a reasonable correlation with future job performance [13], but in dozens of studies has been found to be vulnerable to response distortion when used for higher stakes purposes (selection to professional schools) [14]. Numerous tools have been created to combat response distortion (e.g., Balanced Inventory of Desired Responding), with limited success [15]. However, seemingly innocent adjustments in test format can have dramatic impact on the level of response distortion. For example, in a situational judgment test, asking test-takers "what would you do" rather than "what should be done" will increase levels of response distortion [16]. In contrast, when responses do not have a clear right or wrong answer, this can help reduce the effects of response distortion [17].

2) Cheating: Cheating is a widespread phenomenon across university campuses, and is more likely for higher stakes examinations, and lowered threat of being caught [18]. This is of particular concern for unproctored online tests, though research has surprisingly shown that cheating is more prevalent in live courses over online courses [19]. Additionally, there are numerous steps that can be implemented to discourage cheating in an online setting, such as randomizing the sequence of test questions to each individual test taker, revealing the next question only after the response window for the preceding question has closed, and only allowing the test to be completed within a particular time frame [20]. Recent technological advancements have also allowed for more novel and sophisticated ways to detect cheating, such as through the capturing of keystroke signatures, automated face detection, and voice identification. As technology continues to improve, threats of online cheating will continue to recede.

3) Coaching Effect: The rise in popularity of high-stakes testing simultaneously increases the popularity of preparatory guides and courses whose purpose is to inform test takers about test content and provide study materials to optimize their performance [21]. While claims of significant gains accrued through preparatory test programs are likely inflated [22, 23], even small gains could be seen as socially regressive, as more advantaged applicants can avail themselves of such programs to a greater extent than their less advantaged peers as these preparatory tools tend to be associated with significant financial cost. However, certain modifications in the test parameter has been shown to reduce coaching effects. For instance, increasing the difficulty of the test can sometimes negate, or even reverse, coaching effects, as demonstrated in a study of situational judgment test coaching [24]. Coaching also seems to be less effective for tests of personal attributes (like the multiple mini-interview, described below) as compared to cognitive tests [22]).

4) Practice Effect: Practice effect, sometimes also referred to as a retest effect, refers to test score changes after prior exposure to an identical or alternate form of this test [24]. In other words, test-taker scores can increase simply as a result of having previously taken the test before. Most standardized tests of cognitive skills demonstrate higher scores for those taking the test for the second time (SAT study; [22]). The same is not yet clearly true for measures of professional skills, with mixed reports over its existence in the literature [22, 26].

B. Test Parameters: Psychometrics

5) Construct Deficiency: It is important to ensure that a broad range of topics are covered by an assessment to avoid blind spots. For instance, a comprehensive math test which only covers linear algebra would likely be construct deficient, as it does not cover other topics that are pivotal for math abilities, such as geometry and calculus. In a similar vein, no credibility would be attached to an SAT consisting of only a few questions, as intuitively we know that one needs a large number of biopsies of applicant knowledge set to adequately measure their cognitive skills. This is also true for non-cognitive skills, as an individual's ability to communicate, collaborate, or act ethically in one situation does not necessarily translate to other situations. Any one individual will vary greatly in their professional performance between one context and another, yet large stock is often placed on one-on-one or panel interviews with limited data points provided. Correlation between independent interview scores is appallingly low (0.20 on a scale of 0 to 1), raised to low (0.4) with highly standardized questions and professionally trained interviewers [27], but still falling well short of the level of reliability (> 0.90) recommended for career-altering decisions [28].

6) Self-Report Bias: While response distortion is usually thought of as intentional, self-assessment, intended as accurate contributes a more pernicious form of response distortion. In repeated studies, typically 80 - 90% of drivers believe that they are no less skilled than the average (50th percentile) driver on the road [29]. Of even greater concern, the very worst performers are the ones most likely to overestimate their skill level [30]. Yet many assessment tools on offer continue to depend upon self-report accuracy. The accuracy of self-reports can be improved when self-evaluations are specific to a particular domain, objective, and low in complexity [31], but these parameters can be difficult to attain in a high-stakes context.

7) Construct Validity: Construct validity refers to the extent in which an assessment measures the intended construct, and is of the utmost of importance for selection researchers. A "math" test which does not assess math abilities would be futile, regardless of its other psychometric properties. There are multiple different ways in which construct validity can be assessed. The strongest support of construct validity is predictive validity – that the scores on the test is able to predict meaningful future outcome, i.e. to correlate with other measures reflective of the same intended construct. Yet the more common method in which construct validity is assessed is through face validity, or stakeholder perceptions, where test-takers and key stakeholders (program directors, industry experts) assess whether the content of the test is measuring what it is intended to measure. This is done not because face validity is more accurate (it is decidedly not) but because it is more readily accessible. While it might be reasonable to use face validity as a placeholder until correlation with measures of the same construct become available, there must be a plan to determine those correlations over time. For instance, it would be unreasonable for a physics professor to be required to prove that the cumulative physics test at the end of the course will predict student's performance on the job after they graduate, but there should be a plan in place to collect that data to help inform future curricular changes.

C. Environmental Parameters: i.e. Educational / Social / Cultural Setting

8) Subgroup differences: Differences in mean test scores between groups constrain diversity and widening access. For the United States, historically under-represented groups include African-American and Hispanic-Latino, but another group relevant to engineering schools are females [32]. All three groups tend to have minor to moderately lower mean test scores on both science grade point average (GPA) and standardized cognitive tests. Similar group effects are inconsistently seen on some but not all tests of professional skills; on occasion, the male / female group effect has even been reversed, with females outperforming males [33, 34].

9) Stakeholder perceptions: Some selection practices persist in common use despite overwhelming evidence to demonstrate their uselessness, or worse, negative impact. Perhaps the best example of this phenomenon is the widespread use of file review. Compared to simple formulaic (mechanistic) methods, file review (holistic) methods have a century long history of worse predictive validity, from original studies on repeat offenders among parolees [35], through prediction of future academic success [36], to a literature review of 136 independent studies [37].

10) Resource Demands: As noted in the engineering educational literature [38], validated tests of professional skills require resources that all too often outstrip the capacity of the schools and/or applicants. It would not be feasible for an engineering program to administer an assessment which would require the investment of more resources (e.g., financial cost, personnel) than is available to them. It is also important to consider the cost to applicants, as higher financial barriers to entry will detract students from lower-SES backgrounds to apply to the program.

IV. Solutions for Assessment of Professionalism for Engineering

There are numerous ways in which professionalism is assessed at the time of admissions for medical programs – many of which can extend to admissions for engineering schools. Below we highlight a few of the tools which show the most promise for use in engineering admissions.

A. Measures that have Strong Support

1) Multiple Mini-Interview (MMI): Many (typically 6 – 12) short (typically 5 – 8 minute) interviews conducted in a circuit by multiple students and multiple interviews in rotating “speed-dating” fashion avoids the obstacles above with the exception of resource demands [39], a problem which may be reduced, in part, by conducting the live interviews online (skype MMI) [40]. With respect to subgroup differences, some but not all MMIs have demonstrated higher mean test scores for females than males [41]. The strongest support for the use of MMIs come from the wealth of evidence demonstrating their ability to predict both short-term outcomes in medical school (e.g., preclerkship performance [42]) and long-term outcomes (e.g., national licensing exam [43]).

2) Selected-response SJT (SJT-sr): SJTs are typically in the selected-response format, where respondents are presented with a hypothetical scenario and then provided with a list of possible responses, to which the test-takers select the best response, or rate the effectiveness of each response. The history of SJTs have spanned over 50 years, used initially for selecting military personnel, and more recently and more broadly extended to employee selection [44], and finally most recently for school admissions, particularly medical schools (e.g. United Kingdom Clinical Aptitude Test includes an SJT [45]). Multiple studies generally demonstrate moderately strong psychometric properties and mild to moderate correlations with subsequent performance on measures of the same construct (interpersonal course grades [46]; supervisory ratings [45]) and in clinical practice [46]). While SJTs typically tend to incur high costs upfront for test blueprinting and item development, they are usually cost-effective, particularly for SJT-sr, as its scoring is automated so results can be obtained immediately. Of concern, subject matter experts must agree to the correctness of the responses in advance, and their agreement is attainable only when the answers are straightforward [47]. Straightforward for the subject matters experts means straightforward for the test-takers. One study even showed that test-takers can correctly answer SJT-sr questions without having been presented with the scenario [48]. Test-taker scores on SJT-srs in a high-stakes context cluster at the upper spectrum, making it difficult to separate the “excellent” applicant from the “good” applicant [49]. Other concersn exist for selected response SJTs. They are more vulnerable to coaching effects [50], and at times can demonstrate fairly substantial subgroup differences [51].

3) Constructed response Situational Judgment Test (SJT-cr) – SJT-crs have some distinct advantages over SJT – sr (non-straightforward task, so no skewing of scores upward, nor socially regressive coaching effect; smaller or reversed subgroup differences) [50, 52], but as the ratings cannot be automated, its widespread use has occurred only recently with technological advancements brought to bear to limit its otherwise considerable resource requirements. The most popular SJT-cr for medical admissions is CASPer®, which provides test-takers with a situation in which they are told the part they play, following which they have five minutes to type responses to three probing questions relevant to the situation [53]. There are 12 sections of the CASPer® test, 8 video format and 4 written presentation, to which test-takers must respond, and each section’s set of three questions’ responses is evaluated by a different rater to provide 12 independent scores. Unlike the MMI where programs are required to recruit their own raters, the CASPer® test is rated by a centralized pool of raters who are recruited from a diverse set of stakeholders in the profession and in the community. Raters are then subject to training, accreditation, and monitoring of their performance. Taking advantage of technological advances, it is possible to provide test-taker scores to programs less than 3 weeks from test date, for numbers of test-takers exceeding 3,000 per day [54]. Mean test scores are higher for females than males, and group effects disadvantaging African-Americans and Hispanic-Latinos is markedly smaller, though not absent, compared to traditional methods of cognitive assessment [34]. Research to date demonstrates moderate correlation with subsequent performance on measures of the same construct (licensure scores on the personal competency subsections [55].

B. Measures that show Some Potential

1) Structured Interview: Interviews are typically conducted in an unstructured format, where there is little consistency from one candidate to another, from one interviewer to another, and from one setting to another. With so many sources for bias, interview scores are generally highly unreliable. However, when interviews become more structured – by assigning set questions to be asked for every candidate, gathering ratings of candidates from a panel of independent interviewers, using scoring rubrics, standardizing interview training and performance monitoring – their reliability tends to improve [56]. While unstructured interviews are poor predictors of job performance, structured interviews predict for future performance mildly to moderately. Despite the strong support for the MMI, most medical schools have adopted the structured interview, as programs often do not have the capacity to conduct an MMI (e.g., multiple rooms, professional actors, at least 10 interviewers per candidate). Results from an American multi-institutional study shows that subgroup differences on structured interviews tend to mimic those on the MMI [58], and the two scores do tend to correlate with one another [59], though more recent results have tempered expectations, as only the MMI scores, and not the structured interview scores, were predictive of internship performance [60]. Despite the popularity of structured interviews further improvements would be required to commend its use for high stakes decision-making.

2) Standardized Video Interviews: One of the major barriers to in-person interviews is its cost, both for programs and for applicants. Programs are required to find an appropriate space to conduct interviews with hundreds of students, while students are often required to fly around the world to complete their interviews. In an effort to alleviate this burden, organizations have attempted to utilize technology to their advantage. For instance, MMIs have also been attempted to be conducted online, with positive preliminary results on its reliability, but predictive validity for the online version remains unclear [40]. The AAMC has begun pilot testing the Standardized Video Interview for admission into emergency medicine, where applicants are videotaped online, and respond to approximately five questions, each response following a three-minute period for applicants to consider each question. Unlike the skype MMI, the responses are scored asynchronously. However, as the assessment was just recently introduced, it is still in the pilot phase of research and its psychometric properties remains to be seen [60].

3) Assessment Centers: Assessment centers are similar to the MMIs in that potential candidates are asked to attend a central location where everyone will be screened through a battery of assessment tools, including questionnaires, interviews, SJTs, and behavioral stations. This approach provides a more comprehensive assessment of applicants, with assessments typically across a wide set of constructs. Assessment centers have been used in Israel for their medical school admissions, with promising results demonstrating their high reliability [61]. However, their cost tends to be extremely high as multiple assessment methods need to be made available and sufficient space is necessary to deliver these assessments. When applicants from diverse geographic areas are to be assessed, resource demands placed upon applicants and/or schools may become prohibitive.

V. CONCLUSIONS

Many of the challenges faced by postsecondary undergraduate educational engineering programs have been addressed by medical schools with good results, and advances in technology are playing an increasing part of the solutions. Enhanced diversity for gender, race and ethnicity, cheaper test costs for applicants with limited resources, assessment tests of personal competencies with reliability and predictive validity, are all now part of the framework for medical trainee selection, while work is ongoing to create ever more options for the future. For engineering schools, the opportunities are ready and waiting.

REFERENCES

[1] Flexner A, Pritchet H, Henry S. Medical education in the United States and Canada bulletin number four (The Flexner Report). New York: The Carnegie Foundation for the Advancement of Teaching. 1910.

[2] McGaghie WC. Assessing readiness for medical education: Evolution of the medical college admission test. Jama. 2002 Sep 4;288(9):1085-90.

[3] Stokes W. The complaints that reach our grievance committee. Med Ann Dist Columbia 21(3):157-8, 1952.

[4] Papadakis MA, Hodgson CS, Teherani A, Kohatsu ND. Unprofessional behavior in medical school is associated with subsequent disciplinary action by a state medical board. Academic Medicine. 2004 Mar 1;79(3):244-9.

[5] The Ontario doctors' strike. CMAJ: Canadian Medical Association Journal. 1986 Sep 1;135(5):429.

[6] Frank JR, Danoff D. The CanMEDS initiative: Implementing an outcomes-based framework of physician competencies. Medical teacher. 2007 Jan 1;29(7):642-7.

[7] ACGME. https://www.acgme.org/

[8] AAMC. https://www.aamc.org/

[9] United States Department of Labor. Engineering competency model, 2015. http://www.aaes.org/sites/default/files/Engineering%20Competency%20Model_Final_May2015.pdf

[10] Oberst BS, Jones RC. Offshore outsourcing and the dawn of the post-colonial era of Western engineering education. European journal of engineering education. 2006 Jun 1;31(3):303-10.

[11] Alvarez L. 'Dr. Death,' British Serial Killer, Kills Himself, New York Times, 2014 https://www.nytimes.com/2004/01/14/world/dr-death-british-serial-killer-kills-himself.html

[12] Albanese MA, Snow MH, Skochelak SE, Huggett KN, Farrell PM. Assessing personal qualities in medical school admissions. Academic Medicine. 2003 Mar 1;78(3):313-21.

[13] Barrick MR, Mount MK. The big five personality dimensions and job performance: a meta-analysis. Personnel psychology. 1991 Mar;44(1):1-26.

[14] Kreiter CD. A research agenda for establishing the validity of non-academic assessments of medical school applicants. Advances in health sciences education. 2016 Dec 1;21(5):1081-5.

[15] Li A, Bagger J. Using the BIDR to distinguish the effects of impression management and self-deception on the criterion validity of personality measures: A meta-analysis. International Journal of Selection and Assessment. 2006 Jun;14(2):131-41.

[16] Whetzel DL, McDaniel MA. Situational judgment tests: An overview of current research. Human Resource Management Review. 2009 Sep 1;19(3):188-202.

[17] Lievens F, Peeters H, Schollaert E. Situational judgment tests: A review of recent research. Personnel Review. 2008 Jun 6;37(4):426-41.

[18] McCabe DL, Treviño LK, Butterfield KD. Cheating in academic institutions: A decade of research. Ethics &Behavior. 2001 Jul 1;11(3):219-32.

[19] Watson G, Sottile J. Cheating in the digital age: Do students cheat more in online courses?. Online Journal of Distance Learning Administration. 2010;13(1):n1.

[20] Cluskey Jr GR, Ehlen CR, Raiborn MH. Thwarting online exam cheating without proctor supervision. Journal of Academic and Business Ethics. 2011 Jul 1;4(1).

[21] Sackett PR, Schmitt N, Ellingson JE, Kabin MB. High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. American Psychologist. 2001 Apr;56(4):302.

[22] Griffin B, Harding DW, Wilson IG, Yeomans ND. Does practice make perfect? The effect of coaching and retesting on selection tests used for admission to an Australian medical school. Med J Aust. 2008 Sep 1;189(5):270-3.

[23] Kuncel NR, Hezlett SA. Fact and fiction in cognitive ability testing for admissions and hiring decisions. Current Directions in Psychological Science. 2010 Dec;19(6):339-45.

[24] Cullen MJ, Sackett PR, Lievens F. Threats to the operational use of situational judgment tests in the college admission process. International Journal of Selection and Assessment. 2006 Jun;14(2):142-55.

[25] Lievens F, Buyse T, Sackett PR. Retest effects in operational selection settings: Development and test of a framework. Personnel Psychology. 2005 Dec;58(4):981-1007.

[26] Dore K. Online Situational Judgement Tests (CASPer): Implications and perspectives of test security. 2017 Novemberl Presented at the annual Australian Association for Research in Education (AARE) conference.

[27] Kreiter CD, Yin P, Solow C, Brennan RL. Investigating the reliability of the medical school admissions interview. Advances in Health Sciences Education. 2004 Jun 1;9(2):147-59.

[28] Downing SM. Reliability: On the reproducibility of assessment data. Medical education. 2004 Sep;38(9):1006-12.

[29] Svenson O. Are we all less risky and more skillful than our fellow drivers?. Acta psychologica. 1981 Feb 1;47(2):143-8.

[30] Eva KW, Regehr G. Exploring the divergence between self-assessment and self-monitoring. Advances in Health Sciences Education. 2011 Aug 1;16(3):311-29.

[31] Zell E, Krizan Z. Do people have insight into their abilities? A metasynthesis. Perspectives on Psychological Science. 2014 Mar;9(2):111-25.

[32] Hackett G, Betz NE, Casas JM, Rocha-Singh IA. Gender, ethnicity, and social cognitive factors predicting the academic achievement of students in engineering. Journal of counseling Psychology. 1992 Oct;39(4):527.

[33] Reiter H, Eva K. Vive la difference: The freedom and inherent responsibilities when designing and implementing multiple mini-interviews. Academic Medicine. 2018 Jul 1;93(7):969-71.

[34] Juster F, Miller D, Dore K, Reiter HI. Are medical student demographics impacted by implementation of a situational judgment test? Paper presented at Canadian Conference on Medical Education (CCME) in Montreal, Canada, 2015.

[35] Burgess EW: Factors determining success or failure on parole. In Bruce AA, editor: The workings of the indeterminate sentence law and the parole system in Illinois, Springfield IL: Illinois Committee on Indeterminate Sentence Law and Parole, 1968, pp 205-249 (originally published in 1928)

[36] Sarbin TR. A contribution to the study of actuarial and individual methods of prediction. American Journal of Sociology. 1943 Mar 1;48(5):593-602.

[37] Grove WM, Meehl PE: Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: the clinical-statistical controversy. Psychol Public Policy Law 2:293-323, 1996

[38] Prus J, Johnson R. A critical review of student assessment options, in Assessment and Testing: Myths and Realities, Bers TH and Mittler ML eds, New Directions for Community Colleges, San Francisco: Jossey-Bass, No. 88, Winter 1994, pp 69-83

[39] Eva KW, Rosenfeld J, Reiter HI, Norman GR. An admissions OSCE: The multiple mini-interview. Medical education. 2004 Mar;38(3):314-26.

[40] Tiller D, O'mara D, Rothnie I, Dunn S, Lee L, Roberts C. Internet-based multiple mini-interviews for candidate selection for graduate entry programmes. Medical education. 2013 Aug;47(8):801-10.

[41] Eva KW, Reiter HI, Rosenfeld J, Trinh K, Wood TJ, Norman GR. Association between a medical school admission process using the multiple mini-interview and national licensing examination scores. Jama. 2012 Dec 5;308(21):2233-40.

[42] Eva KW, Reiter HI, Rosenfeld J, Norman GR. The ability of the multiple mini-interview to predict preclerkship performance in medical school. Academic medicine. 2004 Oct 1;79(10):S40-2.

[43] Eva KW, Reiter HI, Rosenfeld J, Norman GR. The relationship between interviewers' characteristics and ratings assigned during a multiple mini-interview. Academic Medicine. 2004 Jun 1;79(6):602-9.

[44] Lievens F, Peeters H, Schollaert E. Situational judgment tests: A review of recent research. Personnel Review. 2008 Jun 6;37(4):426-41.

[45] Patterson F, Cousans F, Edwards H, Rosselli A, Nicholson S, Wright B. The predictive validity of a text-based situational judgment test in undergraduate medical and dental school admissions. Academic Medicine. 2017 Sep 1;92(9):1250-3.

[46] Lievens F, Sackett PR. The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. Journal of Applied Psychology. 2012 Mar;97(2):460.

[47] De Leng WE, Stegers-Jager KM, Husbands A, Dowell JS, Born MP, Themmen AP. Scoring method of a situational judgment test: Influence on internal consistency reliability, adverse impact and correlation with personality?. Advances in Health Sciences Education. 2017 May 1;22(2):243-65.

[48] Krumm S, Lievens F, Hüffmeier J, Lipnevich AA, Bendels H, Hertel G. How "situational" is judgment in situational judgment tests?. Journal of Applied Psychology. 2015 Mar;100(2):399.

[49] Harris BH, Walsh JL, Lammy S. UK medical selection: lottery or meritocracy?. Clinical Medicine. 2015 Feb 1;15(1):40-6.

[50] Lievens F, Peeters H. Impact of elaboration on responding to situational judgment test items. International Journal of Selection and Assessment. 2008 Dec;16(4):345-55.

[51] Lievens F, Patterson F, Corstjens J, Martin S, Nicholson S. Widening access in selection using situational judgement tests: evidence from the UKCAT. Medical education. 2016 Jun;50(6):624-36.

[52] Lievens F, Sackett PR. The effects of predictor method factors on selection outcomes: A modular approach to personnel selection procedures. Journal of Applied Psychology. 2017 Jan;102(1):43.

[53] Dore KL, Reiter HI, Eva KW, Krueger S, Scriven E, Siu E, Hilsden S, Thomas J, Norman GR. Extending the interview to all medical school candidates—Computer-Based Multiple Sample Evaluation of Noncognitive Skills (CMSENS). Academic Medicine. 2009 Oct 1;84(10):S9-12.

[54] TakeCASPer FAQ. https://takecasper.com/faq/#toggle-id-16

[55] Dore KL, Reiter HI, Kreuger S, Norman GR. CASPer, an online pre-interview screen for personal/professional characteristics: Prediction of national licensure scores. Advances in Health Sciences Education. 2017 May 1;22(2):327-36.

[56] Wiesner WH, Cronshaw SF. A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the

employment interview. Journal of Occupational Psychology. 1988 Dec;61(4):275-90.

[57] Henderson MC, Kelly CJ, Griffin E, Hall TR, Jerant A, Peterson EM, Rainwater JA, Sousa FJ, Wofsy D, Franks P. Medical school applicant characteristics associated with performance in multiple mini-interviews versus traditional interviews: A multi-institutional study. Academic Medicine. 2018 Jul 1;93(7):1029-34.

[58] Jerant A, Henderson MC, Griffin E, Hall TR, Kelly CJ, Peterson EM, Wofsy D, Franks P. Do multiple mini-interview and traditional interview scores differ in their associations with acceptance offers within and across five California medical schools?. Academic medicine: journal of the Association of American Medical Colleges. 2018 Mar.

[59] Jerant A, Henderson MC, Griffin E, Hall TR, Kelly CJ, Peterson EM, Wofsy D, Tancredi DJ, Sousa FJ, Franks P. Do admissions multiple mini-interview and traditional interview scores predict subsequent academic performance? A Study of Five California Medical Schools. Academic medicine: journal of the Association of American Medical Colleges. 2018 Sep.

[60] Davis JJ. The 2017–2018 Standardized Video Interview: An Ethical Concern. Academic Medicine. 2018 Jan 1;93(1):11.

[61] Ziv A, Rubin O, Moshinsky A, Gafni N, Kotler M, Dagan Y, Lichtenberg D, Mekori YA, Mittelman M. MOR: A simulation-based assessment centre for evaluating the personal and interpersonal qualities of medical school candidates. Medical Education. 2008 Oct;42(10):991-8.