Journal of Educational Evaluation for Health Professions

J Educ Eval Health Prof 2015, 12: 51 • http://dx.doi.org/10.3352/jeehp.2015.12.51

Open Access

eISSN: 1975-5937

RESEARCH ARTICLE

Developing a situational judgment test blueprint for assessing the non-cognitive skills of applicants to the University of Utah School of Medicine, the United States

Jorie M. Colbert-Getz[1]*, Karly Pippitt[2], Benjamin Chan[3]

Departments of [1]Internal Medicine Administration, [2]Family and Preventive Medicine, and [3]Psychiatry and Assistant Dean of Admissions, University of Utah School of Medicine, Salt Lake City, UT, USA

Abstract

Purpose: The situational judgment test (SJT) shows promise for assessing the non-cognitive skills of medical school applicants, but has only been used in Europe. Since the admissions processes and education levels of applicants to medical school are different in the United States and in Europe, it is necessary to obtain validity evidence of the SJT based on a sample of United States applicants. **Methods:** Ninety SJT items were developed and Kane's validity framework was used to create a test blueprint. A total of 489 applicants selected for assessment/interview day at the University of Utah School of Medicine during the 2014-2015 admissions cycle completed one of five SJTs, which assessed professionalism, coping with pressure, communication, patient focus, and teamwork. Item difficulty, each item's discrimination index, internal consistency, and the categorization of items by two experts were used to create the test blueprint. **Results:** The majority of item scores were within an acceptable range of difficulty, as measured by the difficulty index (0.50-0.85) and had fair to good discrimination. However, internal consistency was low for each domain, and 63% of items appeared to assess multiple domains. The concordance of categorization between the two educational experts ranged from 24% to 76% across the five domains. **Conclusion:** The results of this study will help medical school admissions departments determine how to begin constructing a SJT. Further testing with a more representative sample is needed to determine if the SJT is a useful assessment tool for measuring the non-cognitive skills of medical school applicants.

Key Words: *Cognition; Communication; Judgment; School admission criteria; United States*

INTRODUCTION

Today's physician needs a strong foundation of medical knowledge coupled with non-cognitive skills. These skills include effective communication, teamwork, ethical behavior, and displaying the highest level of professionalism when treating patients. The use of the Medical College Admission Test (MCAT) is pervasive in the admissions process of all United States (US) medical schools. However, this test only focuses on medical knowledge, and neglects the aforementioned non-cognitive areas above. Could there be a way to assess non-cognitive domains? For years, admission committees have strived to assess non-cognitive skills in their candidates using standard interview day experiences, but have not identified an ideal assessment tool. The most recently proposed non-cognitive assessment tool is the multiple mini interview (MMI), which has some evidence of validity regarding the interpretation of scores [1]. However, the MMI is time- and resource-intensive and, with almost 50,000 students applying to multiple US medical schools [2], it is not feasible for every school to have all of their candidates complete a MMI. In Europe, medi-

*Corresponding email: jorie.colbert-getz@hsc.utah.edu

Received: August 5, 2015, Accepted: October 27, 2015;
Published: October 31, 2015
This article is available from: http://jeehp.org/

cal schools and residency programs have used the situational judgment test (SJT) to assess the non-cognitive skills of applicants. They recognize that superior medical knowledge is not enough to become a physician, and the SJT likewise focuses on areas such as ethics, communication skills, teamwork, and professionalism. The interpretation of SJT scores has shown promise in terms of validity evidence for medical schools in the United Kingdom and Belgium [3-5]. To date, no study has been published on the use of a SJT in US medical school admissions. Since the admissions processes and education levels of students applying to medical school are different in the US and Europe, validity evidence is needed for the interpretation of SJT scores in US applicants. Thus, the purpose of this study was to develop a SJT blueprint, drawing upon Kane's validity argument framework.

METHODS

Participants

The participants were 489 applicants selected for assessment/interview day at the University of Utah School of Medicine during the 2014-2015 admissions cycle. Selection for assessment/interview day is based on achieving a minimum level of performance in seven areas: undergraduate grade point average, MCAT score, community/volunteer service, leadership, research, physician shadowing, and patient exposure. A total of 41 assessment days were scheduled, and one of five SJTs was used on each assessment day, such that 84-109 applicants completed each SJT. Applicants also completed a MMI and a traditional interview on assessment day.

Situational judgment test

A SJT includes a series of dilemmas. For each dilemma, examinees are asked to rank five responses, from the most appropriate to least appropriate action, or to select the three most appropriate actions (out of eight) to the dilemma. We piloted five SJTs and each test assessed one of five non-cognitive domains: professionalism, coping with pressure, communication, patient focus, and teamwork. These domains are commonly used in European medical school and residency selection decisions. Since this was a pilot, we wanted to see how each domain performed alone with as many items as possible, which is why we did not require applicants to answer questions belonging to multiple domains. The elements of each domain are presented in Appendix 1. Each SJT included nine select-best-three and nine rank-order items. The items were constructed by the Office of Admissions, in conjunction with the Dean of Admissions and members of the Admissions Committee. The overall structure of the SJT was adapted from European reference books [6-8]. An educational expert and admissions committee member who also directs medical school courses proofread all 90 items and five non-medical experts each reviewed 18 items for clarity and comprehensibility.

Two types of SJT items were used (select-best-three and rank-order), and for all items applicants read a dilemma (item stem) with no clear single solution. For select-best-three items, applicants selected the best three options for dealing with the dilemma from a list of eight options. For rank-order items applicants ranked five options based on their appropriateness to the situation, from 1 being the most appropriate to 5 being the least appropriate. Appendix 2 contains an example of each SJT item type. Applicants had 36 minutes (two minutes per item) to complete the SJT. All examinations were administered on iPads using SoftTest. Select-best-three items were worth 12 points total, with 4 points given for each correct option selected. Rank-order items were worth 20 points if all options were in the correct order, 17 points if only the first and last options were in the correct order, 14 points if either the first or last option was in the correct order, and 10 points if the first and last options were off by no more than 3 ranks. For example, if the correct order was ABCDE and an applicant answered ABCDE, he or she would receive 20 points; if an applicant answered ACBDE, he or she would receive 17 points; if an applicant answered BCDAE, he or she would receive 14 points; and if an applicant answered CAEBD, he or she would receive 14 points.

Kane's validity argument framework

We selected Kane's validity argument framework because it allowed us to determine which types of validity were most important and the order in which validity evidence should be collected, which earlier frameworks were not able to do [9]. In Kane's validity argument framework, the first step is to determine the use and interpretation of scores from an assessment. The next step is to derive assumptions from the use and interpretation, much like creating hypotheses. The third step is to test the weakest assumptions of the validity argument by collecting evidence in a step-wise fashion for four areas: (1) scoring, (2) generalization, (3) extrapolation, and (4) implication inferences. Scoring inference deals with the quality of assessment items, raters, and how an actual assessment score is computed. Generalization inference deals with how well the sample of assessment items adequately represents the domain(s) of interest. Extrapolation inference deals with how well the assessment measures real world performance. Implication inference deals with the effect of scores on an assessment (e.g., 70% equals passing) and therefore the consequences of those decisions (e.g., a student with a 68% requires remediation). Depending on the presence of gaps between evidence and assumptions, the assessment may need to be revised or may not be recommended for further use. Thus, it is important to re-

view evidence at each stage in Kane's framework and use that information to drive future evidence collection or to make revisions, rather than gathering all possible evidence at once.

For step one in Kane's validity framework, the use of the SJT is to assess non-cognitive skills in applicants to medical school and the interpretation is that higher scores would indicate better non-cognitive skills. Therefore, the assumption (step two) is that applicants with higher SJT scores would turn out to be better practicing physicians because they have effective interpersonal, communication, teamwork, ethics, and professionalism skills applicable to patient care. We developed a pool of SJT items and conducted this study to gather evidence for scoring inferences and generalization inferences (step three), so that a test blueprint could be constructed for future use in gathering evidence for extrapolation and implication inferences.

Data analysis

All data were analyzed with IBM SPSS ver. 21.0 (IBM Co., Armonk, NY, USA). For scoring inference, we investigated how each domain performed in terms of item difficulty (average point value) and discrimination (corrected point-biserial). Item difficulty ranges from 0 (no one answered the item correctly) to 1.00 (everyone answered the item correctly). An item should have a difficulty index between 0.30 and 0.80, meaning that 30%-80% students answered the item correctly, in order to be considered good [10]. Values below 0.30 suggest the item is too difficult, and values above 0.80 suggest that the item is too easy. In either case, the item would need to be revised for future use. We also determined if any item's difficulty value was two standard deviations above or below the mean. Discrimination indexes range from −1.00 to 1.00. A good discrimination value for corrected point-biserial computation is 0.30 or higher [11]. Values of 0.11-0.29 are fair, while values below 0.11 are poor and suggest that the item should be omitted or is in need of major revision. For generalization inference, we investigated the internal consistency of item scores in each domain with Cronbach's alpha and measured the concordance of two education experts in the categorization of items by each domain. A Cronbach's alpha coefficient of at least 0.70 is considered good [12], while a coefficient below 0.70 suggests that major revision is necessary or that more assessment methods should be used in conjunction with the examination. Percent agreement values were provided as a descriptive measure, so we did not specify a cut-off for what was constituted good concordance between the experts.

Scoring and generalization evidence was used to select the best items from the pool of 90 multiple-domain SJTs. Specifically, items with difficulty values two standard deviations or more above or below the mean were omitted from the question pool. Once the best items were selected and/or constructed (if necessary, based on how many items were omitted), two educational experts categorized the items to determine how well the pre-determined domain matched their categorization and if items assessed multiple domains. The percent agreement was computed between the educational experts for each of the five domains. Specifically, the number of items that the experts agreed upon was summed and divided by the number of all items in each category. The educational experts used the descriptions in Appendix 1 as a rubric for categorization. Items that assessed multiple categories or items that the educational experts could not reach agreement on were revised or omitted from the question pool. Since each item that the experts disagreed on was discussed, we only provided percent agreement as a descriptive measure and did not compute Cohen's kappa. The final goal was to have 60 items for a two-hour test.

Ethical approval

The institutional review board at the University of Utah School of Medicine deemed this study exempt.

RESULTS

The average difficulty and discrimination index for each SJT domain are presented in Table 1. The difficulty index ranged from 8.27 (effective teamwork) to 9.94 (coping with pressure) out of 12 points total for the select-best-three items and 13.72

Table 1. Average item difficulty and item discrimination index of situational judgment test domains based on the responses of 489 applicants to the University of Utah School of Medicine, United States

Item type and item analysis index	Domain				
	Effective communication	Patient focus	Effective teamwork	Commitment to professionalism	Coping with pressure
Select-best-three items					
Difficulty index	0.78	0.78	0.69	0.71	0.83
Discrimination index	0.10	0.04	0.15	0.01	0.07
Rank-order items					
Difficulty index	0.78	0.73	0.77	0.69	0.72
Discrimination index	0.06	0.20	0.15	0.18	0.06

(commitment to professionalism) to 15.55 (effective communication) for the rank-order items. Two rank-order items had a difficulty index two standard deviations below the mean of all rank-order items. One select-best-three item had a difficulty value two standard deviations below the mean, and one select-best-three item (effective communication) had a difficulty value two standard deviations above the mean for all select-best-three items. All four of these items were omitted from the question pool. The average discrimination was fair (0.11 to 0.20) for rank-order items in the coping with pressure, communication, and patient focus domains and also for select-best-three items in the communication domain. All other average discrimination values were poor (< 0.10). Ten rank-order and seven select-best-three items had negative or zero discrimination values, and these 17 items were omitted from the pool.

Cronbach's alpha results for the 69 items (33 rank-order and 36 select-best-three) selected out of 90 items suggested that the items in each domain had low internal consistency (coefficients < 0.60). Additionally, the majority of the items (63%) categorized by the two education experts appeared to be assessing multiple domains (e.g., communication and teamwork). Specifically, 80% of patient focus items, 72% of teamwork items, 71% of communication items, 67% of coping with pressure items, and 33% of professionalism items assessed multiple domains. The concordance of categorization between the two educational experts was higher for the items in the teamwork (72% agreement), professionalism (78% agreement), coping with pressure (76% agreement) and patient focus (70%) domains than in the communication domain (24%).

Final testing blueprint

Based on the generalization inference evidence, we decided to omit the patient focus category from the final testing blueprint. Initially, patient-focused care was felt to be extremely important to the staff and faculty at the University of Utah School of Medicine. However, the dilemmas for these items took place in a hospital environment and may unfairly advantage applicants with prior medical experience. Communication appeared to be present and required for handling many of the SJT dilemmas. Thus, we decided to tighten the description of the communication category to focus specifically on how information should be delivered and received in the context of others. Therefore, we used four domains for the testing blueprint: communication, teamwork, professionalism, and coping with pressure. We did note that rank-order items required more time (and thus cognitive load) for the applicants to complete. These items were also difficult to score with our current testing software. Therefore, we decided to include 10 select-best-three and five rank-order items for each of the four domains, ultimately obtaining a final SJT with 60 items.

DISCUSSION

This is the first study to report the development of a SJT for use in the admissions process to a medical school in the US. Based on Kane's validity argument framework, we were able to create a testing blueprint for future use. The generalization evidence suggested that SJT domains do not have good internal consistency, and two raters found that multiple domains were assessed by many items. Therefore, more generalization evidence is needed for the final 60-item test before moving to collecting extrapolation and implication evidence according to Kane's validity argument framework. Few studies in medical education have used Kane's validity argument framework for developing and validating assessment scores, so this study will provide an example to other health professions educators and administrators.

Several limitations of this study should be considered. First, we sampled a low number of items for each applicant, which was reflected in the low discrimination and reliability values. Since this was a pilot, we wanted to ensure that fatigue did not factor into applicants' selection of answers to the items and also wanted to be able to assess a larger number of items. Second, the SJT questions could have been circulated among applicants, since the assessment dates spanned many months. On the morning of their assessment or interview day, applicants were asked to sign a confidentiality statement in which they agreed not to reveal the SJT scenarios to anyone. The admissions staff monitored known pre-medical student websites and did not see any posts of sensitive SJT material. However, the possibility still exists that this information could have been leaked. Third, this study was conducted among applicants to one medical school. Thus, future validity evidence collection needs to focus on applicants at multiple medical schools. It should also be noted that the majority of the SJT items were found to assess multiple domains and the reliability was low for each domain. However, this finding was based on only two experts' opinions about categorization. Thus, more evidence is needed for generalizability, and more experts may be needed to categorize the items.

The results of this study will help medical school administrators determine how to begin constructing a SJT, and we identified important factors to consider based on Kane's validity argument framework. Specifically, we will continue to build our SJT question bank with more of a focus on dilemmas that assess a single non-cognitive domain. Further testing of the validity argument framework with a more representative sample, perhaps including multiple institutions, is needed to determine if the SJT is the optimal assessment tool for measur-

ing the non-cognitive skills of medical school applicants. The SJT is one of many possible ways to assess non-cognitive skills. Data from such testing will need to be followed to determine whether the SJT has long-term predictive value. If successful, the SJT could be included as a portion of the initial assessment of applicants, in either a testing format such as the MCAT, or perhaps as part of an online application prior to granting an interview.

ORCID: Jorie Colbert-Getz: http://orcid.org/0000-0001-7419-7588; Karly Pippitt: http://orcid.org/0000-0002-0596-5907; Benjamin Chan: http://orcid.org/0000-0002-8269-8048

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article was reported.

ACKNOWLEDGMENTS

We would like to acknowledge the Admissions Office staff for their hard work and diligence in administering the SJT. We would also like to thank Kerri Shaffer Carter for her expertise in categorizing items and Wayne Samuelson for his review of the manuscript.

SUPPLEMENTARY MATERIAL

Audio recording of the abstract.

REFERENCES

1. Reiter HI, Eva KW, Rosenfeld J, Norman GR. Multiple mini-interviews predict clerkship and licensing examination performance. Med Educ. 2007;41:378-384. http://dx.doi.org/10.1111/j.1365-2929.2007.02709.x
2. Association of American Medical Colleges. Data and research [Internet]. Washington (DC): U.S. Medical School Applications and Matriculants by School, State of Legal Residence, and Sex; 2014 [cited 2015 Aug 22]. Available from: https://www.aamc.org/download/321442/data/factstable1.pdf
3. Patterson F, Ashworth V, Zibarras L, Coan P, Kerrin M, O'Neill P. Evaluations of situational judgement tests to assess non-academic attributes in selection. Med Educ. 2012;46:850-868. http://dx.doi.org/10.1111/j.1365-2923.2012.04336.x
4. Lievens F, Sackett PR. The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. J Appl Psychol. 2012;97:460-468. http://dx.doi.org/10.1037/a0025741
5. Patterson F, Baron H, Carr V, Plint S, Lane P. Evaluation of three short-listing methodologies for selection into postgraduate training in general practice. Med Educ. 2009;43:50-57. http://dx.doi.org/10.1111/j.1365-2923.2008.03238.x
6. Mahesan N, Choudhury SM, Rymer J. Get ahead!: the situational judgement test. Boca Raton: Taylor & Frances Group; 2012.
7. Metcalfe D, Dev H. Situational judgement test. 2nd ed. Oxford: Oxford University Press; 2014.
8. Picard O, Allsopp G, Campbell L. Foundation programme: 250 SJTs for foundation year entry. London: ISC Medical; 2012.
9. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. Med Educ. 2015;49:560-575. http://dx.doi.org/10.1111/medu.12678
10. Tavakol M, Dennick R. Post-examination analysis of objective tests. Med Teach. 2011;33:447-458. http://dx.doi.org/10.3109/0142159X.2011.564682
11. McDonald ME. The nurse educator's guide to assessing learning outcomes. 3rd ed. Burlington (MA): Jones & Bartlett Publishers; 2014.
12. Tavakol M, Dennick R. Post-examination interpretation of objective test data: monitoring and improving the quality of high-stakes examinations: AMEE Guide No. 66. Med Teach. 2012;34:e161-e175. http://dx.doi/10.3109/0142159X.2012.651178

Appendix 1. Domains and elements of the situational judgment test for assessing the non-cognitive skills of applicants to the University of Utah School of Medicine, USA

Domain A. Effective communication
- Is reliable and punctual
- Takes responsibility for own work and actions
- Understands the emotion and intent behind the message
- Includes non-verbal cues and body language
- Includes written communication, both traditional and electronic
- Includes active listening
- Ability to assert oneself when situation calls for it

Domain B. Patient focus
- Care that includes respect and responsiveness to the patient, addressing the patient's needs and values
- Takes into account the background (culture) of the patient
- Has patient take an active role in their care

Domain C. Effective teamwork
- Commitment to a goal
- Each team member is heard and makes a valuable contribution
- Participation by all members, be it verbal or with specific tasks
- Decisions are made in a logical manner, and, after internal discussion where dissenting opinions are talked about, individual members support the team decision
- New and innovative ideas have ways to be raised and respected (brainstorming)

Domain D. Commitment to professionalism
- Accountability of one's own actions
- Respect to others, including team members and patients
- Displays high integrity and ethics
- Guards the competency of themselves and other members of the profession
- Maintains the duties and responsibility of the profession

Domain E. Coping with pressure
- Maintains clear and professional communication in the presence of stressful situations
- Continues to display highly integrous and ethical behavior in difficult situations
- Does not escalate situations when others attempt to do so (i.e., does not raise voice or become frustrated when others are doing so)
- Able to maintain appropriate relationships with team members
- Able to keep the focus on the main goal

Appendix 2. Example of situational judgment test items for assessing the non-cognitive skills of applicants to the University of Utah School of Medicine, USA

Select-best-three item

Stem: In your second year of medical school, you have been taking part in three inter-professional simulation exercises with a first year nursing student and third year pharmacy student. The nursing student used to be a medical student (she completed three years of medical school and then dropped out to be a nursing student) and thus is very knowledgeable, but you feel that she can be overbearing about presenting the simulated patient. You feel undermined in your position as the 'doctor' in the simulation exercises.

Instructions

Choose the 'three' most appropriate actions to take in this situation.
Rank in order the following actions in response to this situation from 1 = most appropriate to 5 = least appropriate.

Options

A. Discuss your feelings with the nursing student and ask how she thinks you could overcome this difficulty.
B. Remind the nursing student of your superior position as a doctor in the simulated exercises.
C. Adopt a more confident approach to the simulated exercises.
D. Find an opportunity to challenge the nurse's judgment and demonstrate your superior knowledge.
E. Adopt a more subordinate position as you are less experienced than the nursing student.
F. Speak to the inter-professional course director for advice.
G. Ask the pharmacy student whether he finds the nursing student difficult to work with.
H. Do nothing as long as the nursing student's behavior does not impact your group's performance.

Answer: A, C, F

Rank-order item

Stem: You are in the pediatric clerkship on a service with another medical student. Your fellow medical student has a habit of sending text messages during rounds. The attending has not noticed, but you have seen a number of the children's parents appear less than impressed with the medical student's inattention.

Instructions

Rank in order the following actions in response to this situation from 1 = most appropriate to 5 = least appropriate.

Options

A. Let the medical student know that others have noticed her sending text messages.
B. Ask the medical student if everything is OK.
C. Suggest that the medical student put her phone away immediately.
D. Inform the attending.
E. Create a 'politeness code' including a rule against texting and ask all members of the team to sign the code.

Answer: 1. B, 2. A, 3. C, 4. D, 5. E