Composite Scores and Decision-Making in Undergraduate Medical Education

BY

SOLA AOUN BAHOUS
M.D., Lebanese University, Lebanon, 1995
M.S., Claude Bernard University, France, 2000
Ph.D., Pierre and Marie Curie University Paris VI, France, 2005

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Master in Health Professions Education
in the Graduate College of the
University of Illinois at Chicago, 2017

Chicago, Illinois

Defense Committee:

Ara Tekian, Chair and Advisor
Ilene Harris
Yoon Soo Park

This thesis is dedicated to my husband and friend, Joudy, and my three angels, Jana, Rawane and Jad, for their great love, remarkable understanding and unconditional support, without which it would never have been accomplished.

I dedicate this thesis to my late father as well, who believed in academic excellence and continual learning. His words of wisdom shall not be forgotten.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my mentor, committee chair and advisor, Ara, whose never-ending support and guidance throughout my MHPE journey have shaped-up my personality as educator and scholar. I would also like to thank my thesis committee members, Ilene and Yoon Soo for their unwavering assistance and encouragement. The realization of this work would have never been possible without the diligent supervision and accompaniment of this committee. For all this, and much more, they are wholeheartedly thanked.

I would also like to acknowledge a number of individuals who contributed to data collection and analysis, Rebecca Assaf, Angelique Salloum, Hadil Attieh and Pascale Salameh, and those who participated in the conduct of the study, Vanda Abi Raad, Nazih Youssef, Nadia Asmar, Rania Sakr, Maroun Matar, Michel Samaha and Wael Salameh. Another person who was not only helpful but also always available and caring, Janet Settle, is to be deeply thanked.

Sola

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACGME	Accreditation Council on Graduate Medical Education
ANOVA	Analysis of Variance
CBME	Competency-Based Medical Education
CCC	Clinical Competency Committee
CK	Clinical Knowledge
CPE	Clinical Performance Evaluation
CS	Clinical Skills
GME	Graduate Medical Education
ICC	Intraclass Correlation Coefficient
IM	Internal Medicine
IRB	Institutional Review Board
LAU	Lebanese American University
MCQ	Multiple-Choice Question
Mini-CEX	Mini-Clinical Evaluation Exercise
Ob-Gyn	Obstetrics and Gynecology
OSCE	Objective Structured Clinical Examination
PAC	Performance Appraisal Committee
PC	Primary Care
SPB	Student Promotion Board
UME	Undergraduate Medical Education
USMLE	United States Medical Licensing Examination

SUMMARY

Composite scores are frequently used in medical education to reflect aggregate information about a student's performance. Combining assessment scores into composites has been shown to be a successful practice in traditional medical education models and is normally driven by the educational system in place. The recent paradigm shift to competency-based medical education has been associated with many implications for assessment. A major challenge emerged about the procedure of combining assessment information in competency-based models, and the validity of decision-making based on composites. In this study, we examined validity evidence associated with traditional composite scores and consequential decision-making, and that associated with reformulated composites based on the competency framework. Furthermore, a third decision model about students' academic progress was built from deliberations among education experts. All assessment data about third-year medical students were collected, in addition to scores on International exams and information about residency placement.

Our results showed that the reliability of composite scores is adequate for the scope of their use, irrespective of the medical education system that drove their formulation. However, associations were more meaningful and interpretable in the decision model based on the competency framework, in comparison to the traditional model. The three models yielded an absolute agreement in 67.4% of cases, and a re-classification of students' academic status in the rest. Correlations with external criteria (performance on International exams and residency placement) demonstrated that decisions ensuing from the three models are supported by consequential validity evidence, and that the second model, using competency-guided composite scores, provided a better classification accuracy, especially in the borderline spectrum of performance.

Finally, our findings suggest that the use of composite scores is associated with defensible decisions about student advancement irrespective of the medical education model. However, decision models differ with their ability to address the challenge of identifying struggling students. Although the advancement of competency-based medical education had implications over assessment, formulating composite scores as measures of competencies is feasible and seems to yield better classification decisions.

I. INTRODUCTION

A. Background

For making decisions about students' performance, educators frequently combine scores from multiple tests, yielding one composite score that represents an indicator of a student's overall performance (1). This practice is rooted in the common understanding that combining scores of different but inter-related aspects of a performance provides a panoramic assessment of that performance by comparison with the granular view generated by looking independently at individual components of a performance. Composite scores are used in various domains and accordingly, they might have different meanings. In the food industry for example, composite scores help reducing information overload so that consumers get a fast impression about a product (2). In ranking schools and universities, different measures are combined into an overall score that indicates which school or university performs better than the others (3). In admissions to medical schools, information from various sources is combined to make a selection recommendation about each applicant (1). In medical education, trainees are routinely tested for their abilities in three domains: knowledge, skills and attitudes, and their overall performance frequently is determined from a composite that combines scores obtained in these different domains (4, 5).

The use of composite scores has been frequently challenged in the literature with regard to validity, reliability, and psychometric characteristics of composite components, weighting procedures, and the tendency of composite scores to average performance across domains, with little differentiation or discrimination between trainees (1, 6, 7). Each of these concerns has been addressed to a different extent in the literature so that composite scores used in medical education today are frequently defensible when considering the technical and psychometric standards (8).

The recent paradigm shift from traditional time-based medical education to competency-based medical education (CBME) in graduate and undergraduate programs has been associated with many implications for assessment (9, 10, 11). Moreover, the concept of 'entrustment' that was recently introduced (12), to constitute the basis for decision-making about academic progress of trainees re-challenges the utility of composite scores in informing such decisions. Therefore, the purposes of this study are twofold: 1) to examine validity evidence supporting the use of composite scores in decision-making in traditional medical education models, and 2) to explore the effect of longitudinal monitoring of performance using composites and that of expert deliberations on decision-making.

B. **Computation of Composite Scores**

To generate composite scores, one should first determine the construct that the composite score is intended to reflect. Subsequently, constituting elements or components should be defined, followed by the application of individual weights to each of these components, following a specific rationale. Every step in this process implies thoughtful reflection about evolving matters such as: Is the construct one-dimensional or multidimensional? In the latter case, what are the different dimensions that define the construct of interest? Are all these dimensions important to consider? Are these dimensions distinct, correlated, or interchangeable? How can different scores measuring these dimensions be combined into a single meaningful score? For example, in awarding driving licenses, the construct is the ability to drive with a reasonable level of safety for self and others. This is a multidimensional construct that requires proper knowledge about traffic law, in addition to adequate driving skills demonstrated through performance. Both components are equally important and independently essential for the construct; therefore, measures of knowledge and skills should be looked at distinctly and independently. In other terms, the two measures do not overlap and are not exchangeable; hence, high performance on one test cannot compensate for a low performance on the other, which mandates a non-compensatory approach in decision-making about granting the license and makes deriving a single composite score unreasonable. On the other hand, the clinical performance of a medical student in a clerkship involves various inter-related dimensions of a skill or competence and encompasses the integration of more than one competency. Although these sets of competencies are not exchangeable, there is frequently a content overlap between them, making a compensatory approach to decision-making more reasonable. In other terms, a medical student may score high on a knowledge test and perform less well on a communication skills station in an OSCE (Objective Structured Clinical Examination). A high performance on one exam may compensate for a low performance on another exam. Therefore, a battery of tests measuring different but overlapping dimensions of a performance can be computed into a composite score, constituting an overall measure of a student's clinical performance (1).

C. Psychometric Characteristics of Composite Scores

Assessment is used to serve one or more of the following purposes: 1) to improve instruction (developmental); 2) to assess achievement level and attainment of competencies (making decisions); 3) to identify and screen for student weaknesses (facilitating interventions); 4) to measure outcomes (informing program evaluation); and 5) to predict future performance (predictive). Combining scores from multiple assessments into one composite score is usually linked to decision-making. Given the high-stake nature of the purpose from generating composites, rigorous quality assurance measures should be applied (10). Issues of validity and reliability of composite scores have confronted educators and researchers for decades (1, 13-19). In demonstrating the validity of inferences made from composite scores one would identify an external criterion and base the judgment on how well the composite score relates to this criterion. However, in the health professions education, it is very hard to isolate only one criterion defining good clinical performance; therefore, we frequently use surrogates of this criterion to conduct validity studies. For example, scores on standardized International exams are often used as indicators of performance in these studies. Furthermore, since reliability puts an upper bound on validity, it has been more frequently examined in the literature. It is generally well known that the reliability of longer tests (composites) is higher than that of shorter tests (components) (1, 20).

Weighting of components has been the most challenging step in generating composite scores, especially in the absence of a real external criterion. Several approaches to differential weighting have been examined in the literature:

1. ***Weighting by importance***: In this approach, weights are attributed according to the importance of the content to be tested. Normally this approach drives test development; for instance, if a content area is more important than another in the general domain, more time is given for its teaching and more questions assessing knowledge of this content will be included in the exam, indirectly leading to more weight. Similarly, in a composite including a battery of varied assessments, those that test for the more important domain would be given a higher weight (21). In this case, weighting is determined by experts in the subject matter through a consensus process.
2. ***Applying effective weights***: Using this approach, weights are applied based on the statistical contribution of each component score to the total composite variance. If all components are attributed a nominal weight of one, the component with the largest standard deviation will contribute more to the composite score (22). This approach is very rarely used in current competency-based assessment.
3. ***Reliability weighting***: This approach aims to maximize reliability of the composite score by attributing more weight to more reliable tests. Normally, composite score reliability is higher than the reliability of individual components scores, but this depends on the weighting applied, the

correlation between component scores, and their reliability. However, improving reliability of the composite is not necessarily associated with a better validity (1, 18, 19).

4. ***Validity weighting***: In this approach, weights are attributed to maximize validity. Normally, validity studies involving composite scores correspond to correlation analyses between scores and an external criterion of performance. For example, a clerkship composite score can be correlated to performance on International exams testing for similar constructs (4, 23). Similarly, maximizing validity may lower reliability of the composite (19, 22).

Irrespective of the method used to attribute differential weights to components, an important aspect to consider is the standardization of component scores. In other terms, component scores having the widest distribution (largest standard deviation) contribute more weight to the composite compared to those with narrower distributions (smallest standard deviation). Therefore, transforming raw scores into linear standard scores neutralizes the differential weighting that is implied by component distribution and allows a real application of a desired effective weight to each of these components (20, 24). Generally speaking, composite scores are associated with higher reliability of rating compared to individual component rating, and the validity of inferences is determined by reliability and weighting of components, in addition to their correlation.

D. Composite Scores and Decision-Making

In medical education, composite scores are frequently generated to provide an overall impression about a trainee's performance and to inform decisions related to academic progress. Two approaches to decision-making involving composite scores exist: compensatory and non-compensatory (1). In the compensatory approach, performance on one test compensates for performance on another test and is frequently applied in education, especially when overlap between components of the same construct exists. In the non-compensatory or conjunctive approach, individual components measure different and independent dimensions of the construct, and hence mandate passing each component (or some components) separately in order to pass the whole test. The non-compensatory approach is not frequently applied in medical education, except in a few cases where constructs directly impacting patient safety are involved. For example, while performing a procedure (inserting an intravenous line for example), a trainee may complete all necessary steps impeccably but may omit to correctly identify the patient. Therefore, a procedure that is correctly performed on the wrong patient significantly endangers healthcare. In such instances, although items on the procedure checklist may be weighted equally, the item related to patient safety will be considered separately and a failing grade will be given to the attempt if this step is omitted, even if the checklist score was high. In this case, a blended approach (compensatory and non-compensatory) may apply (25). For example, a trainee ***must*** perform, appropriately, specific items on the checklist (non-compensable items) ***and*** satisfactorily complete a pre-established number of the remaining items (compensable). The non-compensatory approach to decision-making then predominates (20).

E. Competency-Based Medical Education and Composite Scores

In the new era of CBME, curricular planning is conducted backward from pre-established desired outcomes to content design, time in learning is de-emphasized in favor of mastery of learning, and trainees are expected to demonstrate competence before progressing academically (9). Achieving competence frequently involves the integration of various abilities (cognitive, psychomotor and affective); hence, assessment in CBME continues to rely on a thoughtful and meaningful combination of assessment data about a trainee's performance from multiple sources of information (26, 27). In 2010, Holmboe and colleagues identified key features of effective assessment in a competency-based education. These include: 1) continuous and frequent assessment with formative and summative components; 2) criterion-based assessment using a developmental approach; 3) competency-based assessment, including robust work-based assessment tools; 4) assessment tools meeting minimum quality standards; 5) inclusion of a qualitative and narrative approach to assessment; and 6) involvement of various stakeholders in assessment (10). Despite the availability of guiding principles, many challenges persist in designing effective assessments, especially with the more recent shift towards "entrustment" as a potential outcome in CBME and the need to continuously make high-stake decisions. As clearly stated by Holmboe and colleagues (10),

> "At the program level, effective assessment provides the information and judgment necessary to enable program-level decisions about trainee advancement to be made reliably and fairly." (2010, p. 676)

assessment data could be aggregated in different ways to inform decision-making as long as psychometric validity evidence is secured.

In undergraduate medical education (UME), competency framework highlights a number of challenges compared to graduate medical education (GME), where the performance of the graduate is more readily defined by the criteria of the specialty (10). In GME, CBME is implemented using an organizing framework of competencies and assessment data are compiled in a portfolio and analyzed by the Clinical Competency Committee (CCC). In UME on the other hand, applying a competency framework has been more challenging with regard to both, design and assessment (11), and content-specificity seems to impact more significantly decisions about academic progress. More specifically, medical students during their clinical years rotate in different clerkships, which mandates in traditional clerkship models a certain level of achievement in the specialty (clerkship) itself. Decisions about student's advancement to the senior clinical year will then depend on achievements in each clerkship block, normally using composites of clerkship assessment scores. On the other hand, the competency framework implies that decisions on students advancement be made longitudinally across clerkships, based on aggregate assessment data (as distinguished from the clerkship block), either using composites of longitudinal assessment scores or relying on deliberations among educators, similar to those conducted in a CCC at the graduate level. While the use of composite scores in traditional education models has been supported by validity evidence, it is still unclear how to combine scores from different assessments to determine competence and inform decision-making in CBME (11). In the traditional model for example, students rotating in Internal Medicine are tested for their knowledge in the specialty (using a multiple-choice questions exam for example), skills in history taking, physical examination, communication, and others (using OSCE and direct observations), and attitudes (using OSCEs and direct observations), and their scores on these tests are combined to facilitate decisions about passing the clerkship. In the outcome-based model, determining competence in Patient Care for example [one of the Accreditation Council on Graduate Medical Education (ACGME) competencies] relies on a pre-requisite knowledge, and the integration of this knowledge with the required skills and necessary attitudes to deliver optimal patient care. Identifying measures of milestones in Patient Care and combining scores to determine competence remain a real challenge. Using the Albert Einstein quote, “Not everything that counts can be counted and not everything that can be counted counts”, we suggest that the generation of composite scores be based on a clear rationale that brings meaning to assessment in the context of the profession as suggested in a recent study by Park YS et al (5).

At the Lebanese American University (LAU) School of Medicine in Lebanon, the clinical years are designed following the traditional clerkship model, whereby students rotate in clerkship blocks and complete exams specific for each clerkship. At the same time, outcome competencies are defined based on the CanMEDS roles (Appendix A) and students' performance on each of the competencies is monitored throughout the year without impacting any decision toward students' progress. Instead, composite scores generated from clerkship exam grades inform summative decisions. Therefore, the purpose of this thesis is to examine validity evidence supporting the use of composite scores in decision-making in different medical education models and to contrast these decisions with those made from deliberations among education experts.

F. Research Questions

Research question 1: What are the sources of validity evidence associated with the use of traditional composite scores in UME to justify decision-making?

Research question 2: How does the re-computation of composite scores using the competency framework impact decision-making?

Research question 3: How does deliberation among education experts about students' performance impact decision-making?

II. METHODS

A. The Educational Program

The MD program at LAU follows the American model of medical education (four years) and matriculates between 45 and 55 students each year. The program adopts outcome competencies derived from the CanMEDS roles (Appendix A). The clinical years (Med III and IV) include a traditional clerkship model of clinical rotations. The third year consists of seven core clinical clerkships including internal medicine (IM), surgery, pediatrics, obstetrics and gynecology (Ob-Gyn), primary care (PC), neurology, and psychiatry. These clerkships are distributed throughout the academic year and students rotate on different clerkships in pre-established groups. The fourth clinical year consists of more specialized clerkships, selective rotations, and electives. Recommendations for academic progress are made by the Student Promotion Board (SPB), formed of various Deans, Department Chairs and Clerkship Directors. Decisions are made using the Grading and Promotion policies of the school that mandate in the third clinical year passing each of the core clerkships as a condition for promotion to the senior clinical year. Achieving competence is therefore defined for the clerkship and not for the competency domain. The SPB makes four types of decisions consistent with the school policies: 1) promotion to the senior clinical year without any further requirements; 2) promotion to the senior clinical year with academic monitoring (borderline passing students); 3) denial of promotion until a remedial work is completed such as repeat of a clerkship or a course; and 4) repeat of the Med III year. Assessment of learning involves the use of both formative and summative tools. A battery of tests that are specific for each clerkship includes multiple-choice questions (MCQs) and OSCEs, in addition to workplace-based assessment [clinical performance evaluations, mini-clinical evaluation exercises (mini-CEX), and 360 degree patient and staff evaluations]. Formative assessments (mini-CEX and 360 degree evaluations) are used only to provide feedback and do not contribute to decision-making. Scores on summative assessments, on the other hand, (MCQs, OSCEs and clinical performance evaluations) are combined into composites using the weighted means to inform decision-making. Weighting has been determined for each clerkship, based on consensus among content experts, who were guided by general recommendations published by the Dean's Office. These recommendations involve the following: 1) use of the modified Angoff method to set standards for passing MCQs examinations and the Borderline Group method to set standards for passing OSCE; 2) transformation of raw scores into linear standard scores before applying differential weights; and 3) application of differential weights that value clinical evaluations by attributing at least an equal contribution of clinical evaluations as other assessment tools. Standard setting procedures are conducted by trained faculty who are experts in the content area of each MCQs examination and who are direct observers of examinees for OSCE. The modified Angoff is used to set standards of passing for MCQs examinations and the borderline group method for each OSCE station. Passing score for all tests is set arbitrarily at 70% and all scores are scaled to this passing standard after standard setting procedures. Decision-making involves a blended approach whereby passing a clerkship requires a passing clerkship composite score and a passing score on the clinical performance evaluation (CPE). The latter depends on frequent and direct observations of trainees by trained judges who provide rating on 10 items following a developmental scale with clear anchors (Appendix B). Concomitantly, performance of trainees in the different competency domains is monitored but without any impact on decision-making. Students have the option to take the United States Medical Licensing Examinations (USMLE). However, scores of examinees on these exams are never used for promotion purposes; instead, they are considered in internal program evaluation.

B. Conceptual Frameworks

Two complementary conceptual frameworks were selected to situate this study. The first one frames the approach to validity that is adopted in defending the use of composite scores. The second framework relates to decision-making that is used in expert deliberations and judgments about students' academic progress.

1. Validity conceptual framework

The concept of validity has evolved over time. The criterion model was first introduced and consisted in validating an assessment based on how accurately it estimates an external criterion (14, 28). The content model evolved later to attribute validity, based on a sample of performance, only if this sample satisfies specific criteria of representativeness, fairness, and controlling for errors (29, 30). More recently, the construct model was advanced, defining validity as the evidence that supports inferences made from assessment data. Therefore, the contemporary validity theory as described by Messick and Kane (31-33) embraces a unitary approach to validity as *construct* validity, and involves a process of building scientific arguments to support or refute intended interpretations of assessment data. Five sources of validity evidence have been suggested by Messick (1995): *content*, *response process*, *internal structure*, *relations to other variables*, and *consequences of testing* (31). In this study, Messick's validity framework is used to examine existing validity evidence related to *internal structure, relations to other variables* and *consequences* supporting the use of clerkship and end-of-year composite scores in the third medical year. Furthermore, reformulated composites based on the competency framework are tested for *internal structure, relations to other variables* and *consequential* validity evidence as well. Decisions from expert deliberations are assessed with regards to ensuing *consequences*.

2. <u>Decision-making framework</u>

High-stakes decisions are data-driven. This concept of data-driven decision-making has been explored at large in the literature and involves a multi-step process that has been implemented in different situations and educational contexts (34, 35). This process starts by defining what matters first, then identifying appropriate measures of the construct of interest, followed by information generation through the interpretation of collected information. Swang (36) wrote that:

“Data is made up of raw facts, numbers and text and becomes information when it is put into context so that the relationships, between data can be understood. Knowledge occurs when information is combined with experience and judgment to understand the patterns of the information." (2009, p. 108).

Making sense of assessment data corresponds to formulating plausible assumptions and inferences. Decisions follow this semantic interpretation of data and have consequences attached to them (32). The steps that precede decision-making are sometimes consolidated in a policy (or decision rule) that guides final decisions. In other instances, such as the deliberations that occur in a CCC, decisions are not necessarily determined by policies; instead, semantic interpretations and decisions are entangled. Therefore, evaluating a decision-making process involves the evaluation of its consequences. Kane (33) wrote that: "Policies are not true or untrue, accurate or inaccurate. They are effective or ineffective, successful or unsuccessful." (2006, p. 51). Therefore, evaluating the decision-making process based on traditional composite scores and contrasting it with the competency-based generated composite scores and with expert deliberations will be based on comparison of consequential outcomes (academic progress) and their correlation with other outcome measures (such as performance on International exams and placement in residency programs). This concept joins the consequential validity and relations to other variables as a source of validity evidence in Messick's framework.

In parallel, the ACGME framework on the structure, implementation, function, and utility of a well-functioning CCC (37) has been used to guide the formation and operations of the undergraduate Performance Appraisal Committee (PAC) in charge of making decisions on medical students' progress independent of composite scores.

C. Study Design

1. Data

In order to obtain data about residency placement, we included data from two consecutive classes of 2014-2015 and 2015-2016, where assessment forms, tests and policies remained unchanged. Information about students' performance in the third medical year (Med III), including test scores and narrative assessment and comments was collected from the Dean's office. Furthermore, the following data were obtained as well: clerkship composite scores, end-of-year composite scores, conditions of test administration, exam blueprints and content outline, standard setting procedures and score scaling, SPB decisions, and student performance on the USMLE. Students' placement in residency programs was obtained from the exit surveys. The study proposal was determined by the UIC IRB as not involving "human subjects" as defined in 45 Code of Federal Regulations 46.102 (UIC Research Protocol # 2017-0326). It was also exempted from review by the LAU IRB.

2. <u>Current decision model defined by school policies (decision model 1)</u>

The current decision model involves a blended approach to decisions about students' performance. This conjunctive-compensatory rule states that promotion to the senior clinical year is granted only after the following criteria are fulfilled:

a. Passing all clerkships, which corresponds to obtaining a passing clerkship composite score AND a passing CPE score,

b. Satisfactory completion of all assigned remedial work, and

c. Compliance with policies and code of conduct

Students who achieve the listed criteria but receive a clerkship composite score below 73% on two or more clerkships, will require academic monitoring during their senior year. Any student who fails only one clerkship is granted remedial work (repeat of failed clerkship) that should be satisfactorily completed before promotion is authorized. Students who fail two or more clerkships are denied promotion and should repeat the year.

3. <u>Development of decision models 2 and 3: Consensus building</u>

The PAC consisted of eight members and included Med III clerkship directors, dean of education and education experts at the school. The committee was involved in two major tasks: 1) to develop decision model 2 that depends on composite scores, reformulated to measure program competencies; and 2) to make decisions on promotion based on deliberations that are informed by all available quantitative and narrative assessment results.

In line with the ACGME guidelines, committee members were informed of their roles and responsibilities, the purpose of the meetings and types of decisions expected.

The purpose of the first meeting was to accomplish the following:

a. Determine assessments and assessment components that measure each of the competency domains
b. Agree on specific weights yielding one composite score for each competency domain
c. Formulate a decision rule based on generated composite scores

Prior to their first meeting, members received a detailed description of the program outcome competencies and the assessment system at the school, in addition to documents describing school policies that direct decision-making about trainees' academic progress, which helped them develop a common understanding about school proceedings and practices. Members were asked to reflect on the outcome competencies, and to identify, among available assessments, components that measure each of the competencies. During the first meeting, the committee was briefed about the process of reaching consensus, which is based on cooperation, participation, trust and valuing differences. Then, members received a proposal prepared by the Principal Investigator (SAB) addressing the above three charges and were asked to discuss the proposal and present their opinions and concerns.

The proposal was then refined based on members' input and was circulated in its final version for their approval. The decision rule in this second model was the following:

a. A student must receive a passing composite score for each competency domain to be promoted to the senior clinical year
b. A student who fails the 'Physician as Professional' OR the 'Physician as Care Giver' Role is denied promotion
c. A student who fails two or more of the three remaining Roles is denied promotion
d. A student who fails one of the three remaining Roles is granted a remedial
e. A student with as passing score < 73% on any of the competencies requires monitoring during the senior year

To accomplish the second task, the PAC convened again to discuss promotion of students based on assessment data that included narrative information from formative (mini-CEX and 360 degree assessments) and summative (clinical performance evaluation) tests, in addition to all scores (clerkship MCQs, overall OSCE grades and individual station grades, clerkship clinical performance evaluation with scores on individual items on the form). Members received this information 10 days prior to their second meeting and were asked to formulate a rationale for the selection of students with borderline performance, and to consequently identify these students for discussion and dialogue before a decision is made. During the meeting, each member presented his/her individual rationale and the list of selected cases. Cases selected by more than two members were immediately retained for discussion. Then, the preliminary list (generated from common selections by more than two individuals) was presented to committee members who were asked to raise any concern over any additional student they may want to include in the list. Rationales were discussed and all members reached a consensus about a final list of students whose performance was not considered to be meritorious of a straightforward promotion. All meetings were audio-recorded and transcribed verbatim. Using the grounded theory approach (38, 39), proceedings of the second meeting were inductively analyzed independently by two reviewers (SAB and RS) to determine themes and categories within the presented rationale that underlined members' understanding of and beliefs about a borderline student. The two reviewers met to examine their coding and discuss areas of disagreement. They reached a consensus about a final classification scheme, which was reviewed again by a third party, member of the PAC for triangulation. This qualitative approach served to support validity evidence associated with PAC decisions.

D. Validity Study

1. Decision model 1

Validity of the clerkship and end-of-year composite scores was examined in relation to internal structure, relations to other variables and consequences. Internal structure validity evidence was assessed using reliability estimates of each of the composite components and of the composite scores.

Cronbach alpha for the MCQs exams, OSCEs and CPEs was estimated. An exploratory factor analysis was conducted to identify the structure underlying judges rating of clinical performance. The composite reliability was assessed using the stratified alpha coefficient. Relations to other variables was conducted using association studies among various measures of performance to facilitate triangulation. For example, correlation among local test scores was examined and correlation between clerkship grades, end-of-year grade and other outcome measures or indicators of performance, such as students' scores on the USMLE Step 1 and Step 2 Clinical Knowledge (CK) was analyzed as well. Consequential validity evidence was examined by analyzing associations between decision categories and an external criterion, in this case, performance of students on standardized exams (USMLE) and residency placement.

2. <u>Decision model 2</u>

The PAC in its first meeting agreed on the relative contribution of each test or test component to the competency domains using existing information about the program and outcome competencies and assessment tools and forms (Example of an alignment table in Appendix C). This yielded differential weighting that was applied to assessment data. The obtained composite scores were therefore measures of the school CanMEDS Roles and were used to make decisions on academic progress of students according to the newly defined decision rule. The reliability of the obtained composite scores (Internal structure) was determined, in addition to correlations among composites and between composites and external measures of performance (Relations to other variables). Consequential validity evidence was examined by comparing performance of students on USMLE and residency placement between decision categories.

3. __Decision model 3__

During the second PAC meeting, members agreed on cases that should be granted unconditional promotion, and discussed cases that were considered to have a borderline performance using all available assessment data. Results of deliberations yielded a consensus decision-making about these students' academic progress. Accordingly, committee members identified students who they think would be promoted without any specific academic monitoring, those who would be promoted but necessitating specific monitoring (borderline students), those who should repeat a clerkship or a course, and those who should repeat the year. Agreement level between committee members was examined using intraclass correlation coefficient (ICC) as a reliability measure supporting internal structure validity evidence. Consequential validity evidence associated with this model was examined by testing associations between issuing decisions and the identified external criteria (residency placement and students' performance on USMLE). As stated above, proceedings of this meeting were recorded and answers to the open-ended question: "*what was the rationale behind your selection of borderline students*" and ensuing discussions were used in a qualitative analysis. The purpose was to develop an understanding of how education leaders define a borderline performance and whether they share a common mental model in an educational system where all information is shared and policies are communicated and discussed.

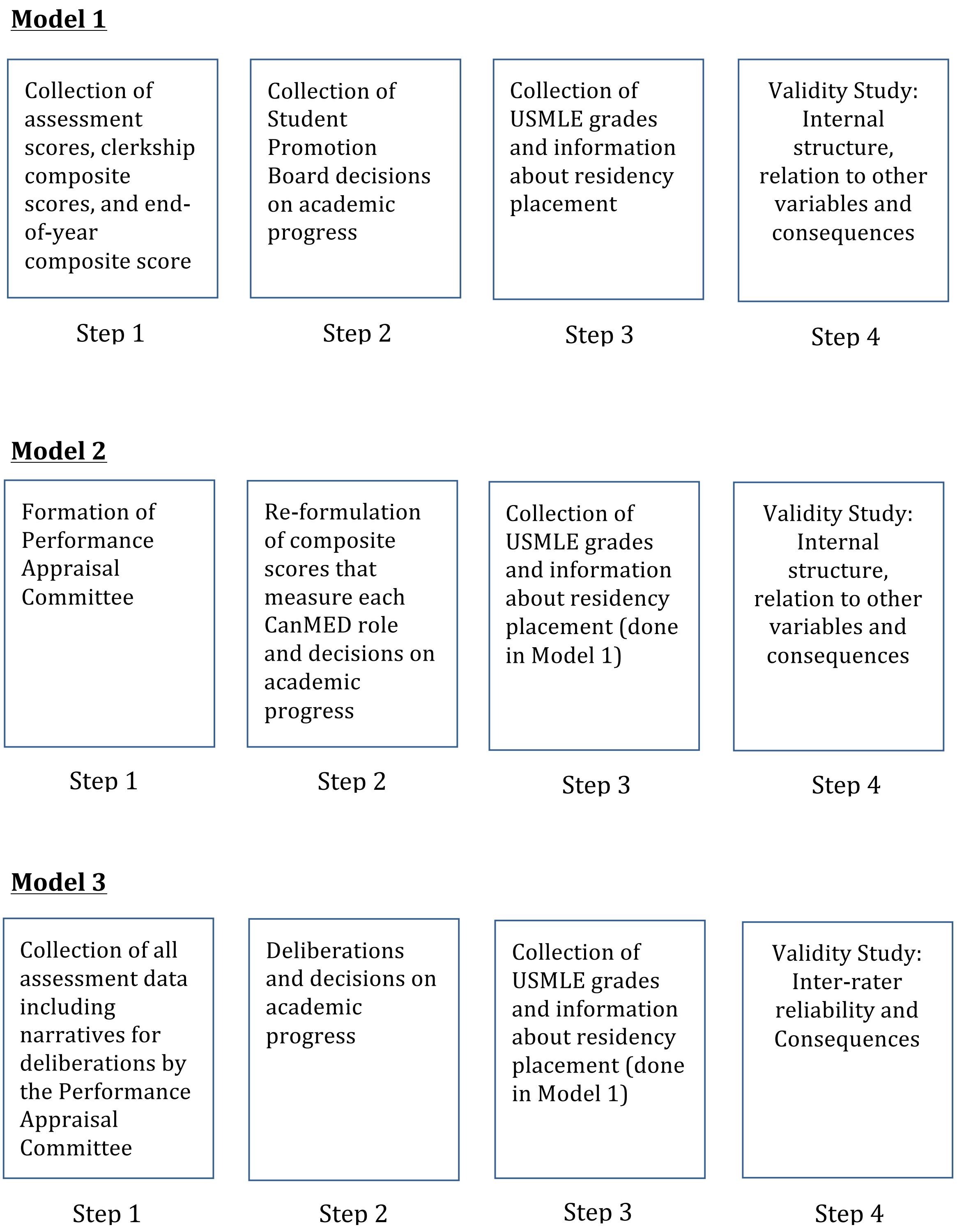
4. Contrasting decision models

The three decision models were contrasted with regards to their associations with USMLE scores and residency placement. Additionally, their level of agreement was tested using Cohen's kappa agreement coefficient, and disagreement was analyzed in relation to the external criteria.

The design of the study is presented in Figure 1.

E. Statistical Analysis

Statistical analysis was conducted on SPSS, version 21.0 for Windows (SPSS Inc., Chicago, USA). Students' component and composite scores, as well as their USMLE scores and residency placement were computed and de-identified at the Dean's Office before data were made available to the research team for statistical analysis. All scores underwent raw-to-scale linear transformation after standard setting procedures were applied. This transformation aimed to have a passing score of 70% for all tests. Residency placement was computed as a three-category discrete variable: matched in top choice, matched in a secondary choice, and unmatched. Descriptive statistics were used to determine means and standard deviations of continuous variables and frequencies and percentages of categorical variables (such as frequencies and percentages of students within each decision category and those within each residency placement category). Cronbach alpha was measured as a reliability estimate for each local test, and stratified alpha for the composite scores, accounting for the reliability of each component score and the attributed weight. An exploratory factor analysis was conducted on ratings of judges provided about students' clinical performance using promax rotation. Kaiser-Meyer-Olkin was used to determine sample adequacy. Factor loading greater than 0.4 was considered significant for retention. Associations between test scores (continuous variables, such as composites and USMLE) were conducted using Spearman rank sum test, and associations between continuous variables and categorical variables (such as between test scores and placement in residency programs) were examined using Kruskal-Wallis test. Chi-squared tests with post hoc analyses using the adjusted standardized residuals were used to compare percentages within and across decision models. The independent-samples *t* test was used for comparison of means between 2 groups. One-way ANOVA was performed to compare means of USMLE scores across decision categories and decision models, and across residency placement categories, where applicable. Post hoc analyses were limited by the small sample size in some groups; therefore, pairwise comparisons were conducted using the independent-samples *t* test. Selection agreement between committee members in model 3 was examined using the ICC. Cohen kappa coefficient was used to determine agreement among different decision models. All tests were double-sided. A two-sided p value < 0.05 was considered statistically significant.



Model 1

Step 1	Step 2	Step 3	Step 4
Collection of assessment scores, clerkship composite scores, and end-of-year composite score	Collection of Student Promotion Board decisions on academic progress	Collection of USMLE grades and information about residency placement	Validity Study: Internal structure, relation to other variables and consequences

Model 2

Step 1	Step 2	Step 3	Step 4
Formation of Performance Appraisal Committee	Re-formulation of composite scores that measure each CanMED role and decisions on academic progress	Collection of USMLE grades and information about residency placement (done in Model 1)	Validity Study: Internal structure, relation to other variables and consequences

Model 3

Step 1	Step 2	Step 3	Step 4
Collection of all assessment data including narratives for deliberations by the Performance Appraisal Committee	Deliberations and decisions on academic progress	Collection of USMLE grades and information about residency placement (done in Model 1)	Validity Study: Inter-rater reliability and Consequences

Figure 1. Study design.

III. RESULTS

A. General Characteristics and Assessment Scores of Study Cohort

Data about 86 Med III students were included in the analysis and are presented in Table I. Mean age was 25.24 ± 0.86 years. Thirty-six students (42%) were females. Average MCQs scores ranged between a minimum of 66.29 ± 6.72 in the Neurology clerkship and a maximum of 79.14 ± 5.65 in the Psychiatry clerkship, OSCE scores between 69.97 ± 9.1 in the Pediatrics clerkship and 84.23 ± 7.37 in the Ob-Gyn clerkship, CPE scores between 82.09 ± 2.56 in the Surgery clerkship and 89.65 ± 4.03 in the Psychiatry clerkship, and clerkship composite scores between 78.18 ± 4.54 in the Neurology clerkship and 84.68 ± 3.28 in the Psychiatry clerkship. Forty-one students (47.7%) presented USMLE Step 1 and scored on average 228.27 ± 17.9 and 33 (38.4%) presented USMLE Step 2 CK and scored on average 237.58 ± 17.26. Twenty-six students (30.2%) presented USMLE Step 2 Clinical Skills (CS) and 25 (96.2%) of them passed.

TABLE I

GENERAL CHARACTERISTICS AND ASSESSMENT SCORES OF STUDY COHORT

Characteristic	**Value (N=86)**
Age mean (SD)	25.24 (0.86)
Gender n (%)	
F	36 (41.9%)
M	50 (58.1%)
Assessment Scores* [Mean (SD)]	
IM	
MCQ (30%)	75.00 (6.96)
OSCE (30%)	79.27 (8.46)
CPE (40%)	83.73 (2.88)
Clerkship Composite Score	79.77 (4.13)
Surgery	
MCQ (30%)	77.43 (5.98)
OSCE (30%)	76.38 (7.78)
CPE (40%)	82.09 (2.56)
Clerkship Composite Score	78.98 (3.33)
Pediatrics	
MCQ (30%)	75.70 (7.77)
OSCE (25%)	69.97 (9.10)
CPE (45%)	85.18 (3.47)
Clerkship Composite Score	78.53 (4.58)
Ob-Gyn	
MCQ (35%)	77.28 (6.29)
OSCE (30%)	84.23 (7.37)
CPE (35%)	84.37 (2.26)
Clerkship Composite Score	81.84 (3.77)
Primary Care	
MCQ (25%)	78.86 (5.95)
OSCE (30%)	74.42 (8.51)
CPE (45%)	82.18 (3.06)
Clerkship Composite Score	78.63 (4.14)
Neurology	
MCQ (30%)	66.29 (6.72)
OSCE (30%)	79.84 (11.9)
CPE (40%)	85.85 (4.00)
Clerkship Composite Score	78.18 (4.54)
Psychiatry	
MCQ (25%)	79.14 (5.65)
OSCE (25%)	80.28 (8.07)
CPE (50%)	89.65 (4.03)
Clerkship Composite Score	84.68 (3.28)
USMLE Grades [Mean (SD) for Step 1 and 2 CK; n (%) for Step 2 CS]	
Step 1 (N=41)	228.27 (17.90)
Step 2 CK (N=33)	237.58 (17.26)
Step 2 CS (N=26)	
Pass	25 (96.2%)
Fail	1 (3.8%)

*Percentages represent percent contribution (weight) of each component score to the clerkship composite score.

Abbreviations: F: Female; M: Male; IM: Internal Medicine; MCQ: Multiple-Choice Questions; OSCE: Objective Structured Clinical Examination; CPE: Clinical Performance Evaluation; Ob-Gyn: Obstetrics and Gynecology; USMLE: United States Medical Licensing Examination; CK: Clinical Knowledge; CS Clinical Skills.

B. Validity Evidence Supporting Decision Model 1

1. Reliability of administered assessments

The reliability of individual assessment tests was estimated using Cronbach alpha and that of composite scores (clerkship composite scores and end-of-year score) using stratified alpha coefficient. Reliability of MCQ scores ranged between 0.449 in Psychiatry and 0.749 in IM, that of OSCE scores between 0.485 in Ob-Gyn and 0.718 in Psychiatry, while all CPE scores had a reliability estimate > 0.8, and clerkship composite scores a reliability between 0.767 in Neurology and 0.888 in Pediatrics. The reliability of the end-of-year composite score was 0.811. Reliability data is presented in Table II.

TABLE II

RELIABILITY ESTIMATES OF ADMINISTERED TESTS

	IM	Surgery	Pediatrics	Ob-Gyn	Primary Care	Neurology	Psychiatry
MCQ	0.749	0.655	0.649	0.675	0.470	0.589	0.449
OSCE	0.534	0.535	0.688	0.485	0.510	0.629	0.718
CPE	0.940	0.910	0.948	0.940	0.888	0.880	0.938
Clerkship Composite Score	0.854	0.774	0.888	0.832	0.800	0.767	0.829
End-of-Year Score	0.811						

Abbreviations: IM: Internal Medicine; Ob-Gyn: Obstetrics and Gynecology; MCQ: Multiple-Choice Questions; OSCE: Objective Structured Clinical Examination; CPE: Clinical Performance Evaluation.

2. <u>Exploratory factor analysis</u>

All correlation coefficients, communalities and anti-image correlations were adequate (all communalities were > 0.25 and factor loadings > 0.4). Kaiser-Meyer-Olkin measure of sampling adequacy was 0.942. Bartlett's test of Sphericity yielded a *P* value < .001. Factor analysis, using the maximum likelihood estimation method with an examination of the Scree plot, could display items over two factors, explaining 63.57% of the total variance of the dataset (49.76% for factor 1, and 13.80% for factor 2). The promax component rotation gave the structure displayed in Table 3. *P* value for all correlation coefficients was <.001. The two factors were labeled as (1) global patient care, and (2) personal and interactive conduct. Cronbach's alpha was 0.883. Results of factor analysis are presented in Table III.

TABLE III

EXPLORATORY FACTOR ANALYSIS: ITEM-LEVEL FACTOR LOADING

Item	Factor 1	Factor 2
Fundamentals of Knowledge	0.813	
Management	0.796	
Critical thinking/Clinical decision making	0.792	
History taking	0.754	
Oral presentations	0.617	
Physical examination	0.608	
Written notes	0.601	
Self reflection		0.892
Professionalism		0.884
Communication skills		0.577
Cronbach alpha (Total: 0.883)	0.877	0.786

3. <u>Relations to other variables</u>

All MCQ exam scores had significant correlation across clerkships (correlation coefficient ranged between r = 0.27, P = .01 and r = 0.6, P < .01). Correlation between OSCE scores across clerkships was significant only for IM with Neurology (r = 0.51, P < .01), IM with PC (r = 0.23, P =.03), IM with Psychiatry (r = 0.37, P < .01), Pediatrics with Ob-Gyn (r = 0.36, P < .01), Ob-Gyn with Neurology (r = 0.29, P < .01), Ob-Gyn with PC (r = 0.25, P = .02), and Neurology with Psychiatry (r = 0.38, P < .01). CPE scores correlated significantly across most clerkships. Associations between tests within a clerkship were very variable. For example, only MCQ and CPE scores were significantly correlated in IM (r = 0.39, P < .01), MCQ and

OSCE scores in Ob-Gyn (r = 0.31, P < .01), MCQ and OSCE scores in Pediatrics (r = 0.33, P < .01), MCQ and CPE scores in Pediatrics (r = 0.23, P = .04), MCQ and CPE scores in PC (r = 0.31, P < .01), and CPE and OSCE scores in PC (r = 0.31, P < .01), while assessment scores did not correlate in Surgery, Neurology and in Psychiatry. USMLE scores correlated significantly with most clerkship MCQ scores, while correlation was non-significant with most of the OSCE and CPE scores (Table IV). Clerkship composite scores correlated significantly across most clerkships except for Surgery with Neurology and Surgery with Psychiatry. Correlation between USMLE Step 1 scores and clerkship composite scores was significant for IM (r = 0.58, P < .01), Pediatrics (r = 0.4, P < .01), and Ob-Gyn (r = 0.53, P < .01), and between USMLE Step 2 CK scores and clerkship composite scores was significant for IM (r =0.54, P < .01), Surgery (r = 0.43, P = .01), Pediatrics (r = 0.55, P < .01), Ob-Gyn (r = 0.8, P < .01), and PC (r = 0.46, P < .01). Given that only one student failed the USMLE Step 2 CS, association studies with this test were not meaningful (Table V).

TABLE IV

CORRELATION MATRIX BETWEEN CLERKSHIP COMPONENT ASSESSMENT SCORES AND USMLE SCORES

	IM MCQ	Su MCQ	Ped MCQ	ObG MCQ	Neu MCQ	PC MCQ	Psy MCQ	IM OSCE	Su OSCE	Ped OSCE	ObG OSCE	Neu OSCE	PC OSCE	Psy OSCE	IM CPE	Su CPE	Ped CPE	ObG CPE	Neu CPE	PC CPE	Psy CPE
IM-MCQ																					
Su-MCQ	0.44**																				
Ped-MCQ	0.37**	0.39**																			
ObG-MCQ	0.58**	0.39**	0.59**																		
Neu-MCQ	0.60**	0.6**	0.23*	0.38**																	
PC-MCQ	0.53**	0.43**	0.28*	0.38**	0.51**																
Psy-MCQ	0.27*	0.31**	0.36**	0.35**	0.27*	0.31**															
IM-OSCE	0.2	0.25*	0.18	0.04	0.18	-0.09	0.15														
Su-OSCE	0.19	0.13	0.09	0.13	0.18	0.33**	0.13	-0.23													
Ped-OSCE	0.38**	0.23*	0.33**	0.18	0.12	0.31**	0.18	0.14	0.12												
ObG-OSCE	0.32**	0.04	0.37**	0.31**	-0.03	0.13	0.25*	0.16	0.06	0.36**											
Neu-OSCE	0.04	0.15	0.42**	0.09	-0.01	-0.21	0.12	0.51**	-0.34**	0.16	0.29**										
PC-OSCE	0.2	0.12	0.39**	0.22*	0.05	0.13	0.14	0.23*	0.05	0.19	0.25*	0.19									
Psy-OSCE	0.13	0.20	0.15	0.15	0.14	-0.03	0.02	0.37**	-0.31**	0.08	0.18	0.38**	0.17								
IM-CPE	0.39**	0.18	0.46**	0.43**	0.11	0.29**	0.28**	0.17	0.17	0.33**	0.35**	0.12	0.38**	0.15							
Su-CPE	-0.05	0.16	-0.06	-0.06	-0.04	-0.06	-0.11	0.28**	-0.06	0.14	-0.08	0.14	-0.08	0.21*	0.23*						
Ped-CPE	0.16	0.23*	0.23*	0.11	-0.05	-0.03	0.04	0.38**	-0.14	0.39**	0.13	0.39**	0.2	0.39**	0.41**	0.55**					
ObG-CPE	0.36**	0.16	0.10	0.21	0.35**	0.53**	0.18	0.05	0.35**	0.22*	0.05	-0.27*	0.28**	0.06	0.43**	0.18	0.22*				
Neu-CPE	0.27*	0.16	0.03	-0.12	0.16	0.31**	0.13	0.12	-0.07	0.29**	0.04	0.03	0.1	-0.01	0.17	0.26*	0.29**	0.28*			
PC-CPE	0.32**	0.11	0.5**	0.32**	0.04	0.31**	0.22*	0.17	0.11	0.43**	0.23*	0.11	0.31**	0.03	0.63**	0.15	0.37**	0.39**	0.21		
Psy-CPE	0.14	0.1	-0.10	-0.14	0.01	0.05	0.02	0.37**	0.06	0.05	-0.08	-0.09	-0.05	0.05	0.24*	0.42**	0.38**	0.28**	0.34**	0.33**	
USMLE Step 1	0.7**	0.26	0.37*	0.60**	0.47**	0.43**	0.3	0.15	0.2	0.24	0.14	0.01	0.18	0.11	0.38*	-0.06	0.14	0.39*	0.05	0.15	0.00
USMLE Step 2 CK	0.69**	0.39*	0.53**	0.78**	0.43*	0.55**	0.20	0.01	0.36*	0.33	0.56**	0.1	0.30	0.12	0.56**	-0.12	0.23	0.32	-0.1	0.31	-0.06

* Correlation is significant at .05 (2-tailed).

** Correlation is significant at .01 (2-tailed).

Light-shaded areas represent correlations between similar assessments across clerkships, while dark-shaded cells represent correlations between different assessments within clerkships.
Abbreviations: IM: Internal Medicine; Su: Surgery; Ped: Pediatrics; ObG: Obstetrics and Gyncology; Neu: Neurology; PC: Primary Care; Psy: Psychiatry; MCQ: Multiple-Choice Questions; OSCE: Objective Structured Clinical Examination; CPE: Clinical Performance Evaluation; USMLE: United States Medical Licensing Examination; CK: Clinical Knowledge.

TABLE V

CORRELATION MATRIX BETWEEN CLERKSHIP COMPOSITE SCORES AND USMLE SCORES

	IM-CSc	Su-CSc	Ped-CSc	ObG-CSc	Neu-CSc	PC-CSc	Psy-CSc
IM-CSc							
Su-CSc	**0.29****						
Ped-CSc	**0.55****	**0.32****					
ObG-CSc	**0.53****	**0.27***	**0.54****				
Neu-CSc	**0.58****	0.14	**0.5****	**0.25***			
PC-CSc	**0.48****	**0.27***	**0.54****	**0.51****	**0.26***		
Psy-CSc	**0.5****	0.21	**0.34****	**0.25***	**0.36****	**0.24***	
USMLE Step 1	**0.58****	0.26	**0.40****	**0.53****	0.25	0.31	0.19
USMLE Step 2 CK	**0.54****	**0.43***	**0.55****	**0.80****	0.23	**0.46****	0.16

* Correlation is significant at .05 (2-tailed).

** Correlation is significant at .01 (2-tailed).

Abbreviations: IM: Internal Medicine; CSc: Composite Score; Su: Surgery; Ped: Pediatrics; ObG: Obstetrics and Gyncology; Neu: Neurology; PC: Primary Care; Psy: Psychiatry; USMLE: United States Medical Licensing Examination; CK: Clinical Knowledge.

4. Consequential validity: Decision model statistics and outcome measures

In this decision model, 76 students (88.4%) were promoted, 7 (8.1%) were offered promotion with academic monitoring, and 3 (3.5%) were required to complete remedial requirements before promotion is considered (Table VI). One-way ANOVA study showed that USMLE Step 2 CK scores differed significantly between decision categories (F (1,31) = 5.061, P = .03), with higher scores noted among promoted students, while differences in USMLE Step 1 scores were non-significant (only one student had completed USMLE Step 1). Among promoted students, 59 (77.6%) matched in top residency choice, 15 (19.7%) matched in their second choice, and 2 (2.6%) did not match. Post hoc analyses using adjusted standardized residuals showed that students who were promoted in this model were more likely to match in first residency choice ($\chi^2_{(2)}$ = 19.4, P < .001). Moreover, students who were offered promotion with academic monitoring had similar chances to match in top residency choice or to fail matching. Students who were requested to complete remedial work either matched in second choice (66.7%) or did not match (33.3%). Students who were offered unconditional promotion had more chances to match in their top residency choice compared to other decision categories ($\chi^2_{(2)}$ = 9.89, P = .007).

TABLE VI

DECISION MODEL 1 AND OUTCOME MEASURES

	Decision Model 1			
	Promotion	Borderline Performance	Remedial	Failure
N(%)	76 (88.4%)	7 (8.1%)	3 (3.5%)	0
USMLE Step 1 [Mean ± SD] (n=41)	228.4 ± 18.1	221[b]	-	-
USMLE Step 2 CK [Mean ± SD] (n=33)	239.2 ± 16.3[a]	212.5 ± 16.3	-	-
Residency Placement [n(%)]				-
Matched in Top Choice	59[c] (77.6%)	4 (57.1%)	0	
Matched in second choice	15 (19.7%)	0	2 (66.7%)	
Unmatched	2 (2.6%)	3 (42.9%)	1 (33.3%)	

[a] $P < .05$.

[b] SD is missing because only one student took the USMLE in this group.

[c] Comparisons between residency placement categories within decision column 1 (Promotion) was significant at the level < .001. Other column proportions did not differ significantly from each other at the .05 level.

Abbreviations: USMLE: United States Medical Licensing Examination; CK: Clinical Knowledge.

C. Validity Evidence Supporting Decision Model 2

1. Reliability of competency-based composite scores

The reliability of competency-based composite scores estimated using the stratified coefficient alpha varied between 0.755 for the ‘Physician as Advocate’ composite score and 0.951 for the ‘Physician as Care Giver’ composite score (Table VII).

TABLE VII

RELIABILITY OF COMPETENCY-BASED COMPOSITE SCORES

CanMED Role	Stratified coefficient alpha
Physician as Scientist	0.838
Physician as Communicator	0.932
Physician as Professional	0.793
Physician as Care Giver	0.951
Physician as Advocate	0.755

2. <u>Relations to other variables</u>

The ‘Physician as Care Giver’ and ‘Physician as Advocate’ composite scores correlated significantly with all other CanMEDS Roles scores, and ‘Physician as Professional’ and ‘Physician as Communicator’ scores correlated significantly with all other scores except for the ‘Physician as Scientist’ composite score. Correlation between USMLE Step 1 and USMLE Step 2 CK scores with composite scores were highly significant for the ‘Physician as Scientist’ score (r = 0.62, P < .01 and r = 0.72, P < .01, respectively) and moderately significant for ‘Physician as Care Giver’ score (r = 0.36, P = .02 and r = 0.52, P < .01, respectively) (Table VIII).

TABLE VIII

CORRELATION MATRIX BETWEEN COMPETENCY-BASED COMPOSITE SCORES AND USMLE SCORES

	Physician as Scientist	Physician as Communicator	Physician as Professional	Physician as Care Giver	Physician as Advocate
Physician as Scientist					
Physician as Communicator	0.172				
Physician as Professional	0.2	**0.32****			
Physician as Care Giver	**0.6****	**0.44****	**0.53****		
Physician as Advocate	**0.46****	**0.33****	**0.71****	**0.66****	
USMLE Step 1	**0.62****	0.07	0.02	**0.36***	0.25
USMLE Step 2 CK	**0.72****	0.2	0.11	**0.52****	0.26

* Correlation is significant at .05 (2-tailed).
** Correlation is significant at .01 (2-tailed).
Abbreviations: USMLE: United States Medical Licensing Examination; CK: Clinical Knowledge.

3. Consequential validity: Decision model statistics and outcome measures

In this decision model, 64 students (74.4%) were promoted, 16 (18.6%) were offered promotion with academic monitoring, 4 (4.7%) were required to complete remedial requirements before promotion is considered, and 2 (2.3%) were denied promotion. One-way ANOVA study showed that USMLE Step 1 and Step 2 CK scores differed significantly between decision categories (F (2,38) = 4.751, P = .01 and F (2,30) = 16.535, P < .01, respectively), with higher scores noted among promoted students. Fifty-eight out of 64 promoted students (90.6%) matched in top residency choice, 5 (7.8%) matched in their second choice, and 1 (1.6%) did not match. Post hoc analyses showed that students who were promoted in this model were more likely to match in first residency choice ($\chi^2_{(2)}$ = 38.90, P < .001).

Moreover, students who were offered promotion with academic monitoring were more likely to match in second choice ($\chi^2_{(2)} = 23.39$, $P < .001$). One out of 4 (25%) students who were requested to complete remedial work matched in top choice, 2 (50%) matched in second choice, and 1 (25%) did not match. One failing student (50%) matched in second choice and the other one (50%) did not match. Students who were offered unconditional promotion had significantly more chances to match in their top residency choice compared to other decision categories ($\chi^2_{(3)} = 39.11$, $P < .001$) (Table IX).

TABLE IX

DECISION MODEL 2 AND OUTCOME MEASURES

	Decision Model 2			
	Promotion	Borderline Performance	Remedial	Failure
N(%)	64 (74.4%)	16 (18.6%)	4 (4.7%)	2 (2.3%)
USMLE Step 1 [Mean ± SD] (n=41)	231.1 ± 15.9[a]	214.6 ± 22	-	-
USMLE Step 2 CK [Mean ± SD] (n=33)	243.3 ± 13.6[a]	216.3 ± 12.4	-	-
Residency Placement [n(%)]				
Matched in Top Choice	58[b] (90.6%)	4 (25.0%)	1 (25%)	0
Matched in second choice	5 (7.8%)	9[b] (56.2%)	2 (50.0%)	1 (50%)
Unmatched	1 (1.6%)	3 (18.8%)	1 (25.0%)	1 (50%)

[a] $P \leq .01$.

[b] Comparison between residency placement categories within decision column 1 (Promotion) was significant at the level < .001. Comparison between residency placement categories within decision column 2 (Borderline Promotion) was significant at the level .001. Other column proportions did not differ significantly from each other at the .05 level.

Abbreviations: USMLE: United States Medical Licensing Examination; CK: Clinical Knowledge.

D. Validity Evidence Supporting Decision Model 3

1. Inter-rater reliability

Eight educators independently identified students whose performance is believed not to be meritorious of a straightforward promotion. The ICC was used to determine agreement level among these educators. Results showed that the ICC was 0.875 (95% Confidence Interval: 0.830 - 0.911). This value represents an indicator of a very good level of agreement among educators (40).

2. Qualitative analysis of committee deliberations

Qualitative analysis of passages from audio-recorded meeting proceedings, that represented discussions of rationales for selection of students, generated 3 major themes that were labeled as follows: 1) Selection guided by valuing; 2) Selection guided by a technical process; and 3) Selection guided by validity. While all members adopted more than one approach for final selection, one approach was considered most important and decisive. Quotations are used to illustrate members' rationale for selection, and some of them fit into more than one theme. Themes and categories are presented in Appendix D.

a. Selection guided by valuing

Some committee members initiated the selection process by identifying values that they believe define a good physician. These values were grouped into 2 categories: *professionalism* and *aggregated skills.* Under these two categories, members selected students by examining comments first, identifying patterns based on repetitive occurrence of the same comment (a qualitative study), and then by observing test scores. Illustrative comments include:

> *"I looked especially at comments to identify students with problem in professionalism. I thought that professionalism cannot be rated well in the others ways."*

> *"I dissected the OSCE stations, some students had 6 times negative comments and low station grades."*

> *"I looked at each OSCE and checked every station separately to identify areas of weakness according to comments then to grades. I thought clinical skills are very important to look at first."*

b. Selection guided by a technical process

Some members defined a technical process to guide their selection. Under this theme, three categories were identified: zoom-in process, pattern recognition and probability building. Two members considered first the overall performance by examining the end-of-year scores. Low performing students necessitated zooming in to focus on detailed performance and deciding on selection. Illustrative comments include:

> *"I started by looking at the final year grade. Then I thought that students who have scores between 70 & 75 are the students I need to know why they have low grades and then I went back to the rotations."*

> *"The end-of-year grade was a first step towards further consideration. Low performers were looked at separately and in details to make my selection."*

Another member considered all scores as a snapshot and used a pattern recognition process to identify students who necessitate further consideration. After those were identified, a final selection was made based on narrative assessment. Illustrative comment includes:

> *"I looked at each final clerkship score for each student and I highlighted the weak students based on my previous experience. Then I read the comments about the highlighted students to select finally those with the bad/negative comments."*

Two members elaborated formulas for selection based on probability and passing scores defined by school policies. Illustrative comments include:

> *"Since I am number-oriented person, I took the cut off value decided by the school of 70/100. So I did a formula to identify how many students have lowest than 70, then I highlighted those cases then I looked horizontally across all the grades. If I had the same pattern across clerkships, then I checked the comments."*

> *"I used the cut off of 70%. Those scoring above 70 on the final clerkship grades were considered eligible for promotion if they had succeeded in the OSCEs. Accordingly, I worked on a formula accounting for most scores but in a rank order. After that I looked at comments. Using the formula, I determined who would*

necessitate discussion because of a high probability of having borderline performance."

c. Selection guided by validity

Some committee members selected students based on assessments that they consider are most accurate. They furthered their selection by looking at narrative assessment. Illustrative comments include:

"…there are too many discrepancies between the CPE and the comments. How can a student receive grades of 8-9 over 10 on an item and then get negative comments. So I did not depend on the CPE scores nor on the comments because I considered them inaccurate."

"…as I told you before the CPE did not reflect accurately students' performance and was not discriminatory (many items are rated all 8 and the narrative did not correlate with the grades) so I went to the OSCE, because I thought it was the most discriminatory."

3. Consequential validity: Decision model statistics and outcome measures

In this decision model, 73 students (84.9%) were promoted, 10 (11.6%) were offered promotion with academic monitoring, 2 (2.3%) were required to complete remedial requirements before promotion is considered, and 1 (1.2%) was denied promotion. One-way ANOVA study showed that USMLE Step 1 scores differed significantly between decision categories while differences in USMLE Step 2 CK scores were non-significant; however, only one student had completed USMLE Step 1 and one completed the USMLE Step 2 CK, which makes comparisons unreliable. Fifty-seven out of 73 promoted students (78.1%) matched in top residency choice, 13 (17.8%) matched in their second choice, and 3 (4.1%) did not match. Post hoc analyses showed that students who were promoted in this model were more likely to match in first residency choice ($\chi^2_{(2)} = 8.16$, $P = .02$). Moreover, students who were offered promotion with academic monitoring had similar chances to either match in top residency choice, match in second choice or to fail matching. One out of 2 (50%) students who were requested to complete remedial work matched in second choice and 1 (50%) did not match. One failing student (100%) did not match. Students who were offered unconditional promotion had significantly more chances to match in their top residency choice compared to other decision categories ($\chi^2_{(3)} = 9.98$, $P < .02$) (Table X).

TABLE X

DECISION MODEL 3 AND OUTCOME MEASURES

	Decision Model 3			
	Promotion	Borderline Performance	Remedial	Failure
N(%)	73 (84.9%)	10 (11.6%)	2 (2.3%)	1 (1.2%)
USMLE Step 1 [Mean ± SD] (n=41)	229.7 ± 15.8	-	173[a]	-
USMLE Step 2 CK [Mean ± SD] (n=33)	238.4 ± 16.8	-	210[a]	-
Residency Placement [n(%)]				
Matched in Top Choice	57[b] (78.1%)	6 (60%)	0	0
Matched in second choice	13 (17.8%)	3 (30%)	1 (50%)	0
Unmatched	3 (4.1%)	1 (10%)	1 (50%)	1 (100%)

[a] SD is missing because only one student took the USMLE in this group.

[b] Comparisons between residency placement categories within decision column 1 (Promotion) was significant at the level < .05. Other column proportions did not differ significantly from each other at the .05 level.

Abbreviations: USMLE: United States Medical Licensing Examination; CK: Clinical Knowledge.

E. Agreement Between Decision Models and Disagreement Analysis

Cohen kappa coefficient was used to determine agreement level between decision models. Results revealed a non-significant agreement between decision models 1 and 2 (kappa coefficient 0.14, *P* = .07), moderate and significant agreement between decision models 1 and 3 (kappa coefficient 0.5, *P* < .001), and low but significant agreement between decision models 2 and 3 (kappa value 0.23, *P* = .006). Absolute agreement was noted in 58 cases (percent agreement 67.4%), 97% of whom were agreement on promotion (Table XI).

Cases where disagreement between models existed were divided into two groups according to which model down-classified the students. For example, disagreements between models 1 and 2 were divided into a group where model 1 attributed a lower promotion (model 1 re-classified the students into lower promotion category) and another group where model 2 attributed a lower promotion classification (model 2 re-classified the students into lower promotion category). Comparison among models showed that model 2 more frequently down-classified students by comparison to model 1 [20 (23.3%) vs. 4 (4.7%), $P < .001$], and to model 3 [17 (19.8%) vs. 6 (7%), $P = .01$]. Moreover, models 1 and 3 down-classified students to the same extent [4 (4.7%) vs. 4 (4.7%), $P = .35$].

TABLE XI

COHEN KAPPA COEFFICIENT AS A MEASURE OF AGREEMENT BETWEEN DECISION MODELS

	Decision Model 1	Decision Model 2
Decision Model 2	0.14	
Decision Model 3	**0.5****	**0.23***

* Significant at < .01.
** Significant at < .001.

1. <u>Comparison between models 1 and 2</u>

Comparison of USMLE scores across agreement/disagreement groups using ANOVA showed significant difference only for USMLE Step 2 CK ($F_{(2,29)} = 9.882$, $P = .001$). Pair-wise comparison using the independent-samples *t* test (post hoc analyses were not possible with the small sample size noted within some groups) showed that, on average, students who were down-classified in model 2 had significantly lower Step 1 USMLE scores by comparison to those who were granted promotion in both models (equal-variance $t_{(38)}$ = 2.36, P = .02), and significantly lower Step 2 CK USMLE scores (equal-variance $t_{(29)}$ = 4.29, P < .0001).

Overall, there was a significant difference in residency placement percentages across agreement/disagreement groups (($\chi^2_{(4)}$ = 34.205, P < .0001). Post hoc analyses using the adjusted standardized residuals showed that students who were down-classified in model 2 were less likely to match in first choice than to match in second choice or not to match ($\chi^2_{(2)}$ = 37.22, P < .0001), and they were less likely to match in first choice by comparison to students who were promoted in both models and those who were down-classified in model 1 ($\chi^2_{(2)}$ = 31.34, P < .0001). By comparison, students who were down-classified in model 1 had similar chances to either match in top choice or not to match (Table XII).

2. <u>Comparison between models 1 and 3</u>

Comparison of USMLE scores between agreement/disagreement groups was limited by the small sample size in the groups. Concerning residency placement, there was an overall significant difference in residency placement percentages across agreement/disagreement groups ($\chi^2_{(4)}$ = 20.118, P < .0001). Post hoc analyses showed that students who were granted promotion in both models had significantly higher chances of matching in top residency choice compared to those in both demoted groups ($\chi^2_{(2)} = 9.18$, $P = .01$). Other comparisons were limited by the small sample size (Table XIII).

TABLE XII

COMPARISON BETWEEN MODELS 1 AND 2: USMLE SCORES AND RESIDENCY PLACEMENT AMONG DIFFERENT AGREEMENT/DISAGREEMENT GROUPS

			Disagreement	
	Total agreement (n=62)	Agreement on Promotion (n=60)	Model 2 down-classified cases (n=20)	Model 1 down-classified cases (n=4)
USMLE Step 1 (mean ± SD)	231.4 ± 16.0 (n=33)	231.4 ± 16.0[a] (n=33)	214.6 ± 22 (n=7)	221[c] (n=1)
USMLE Step 2 CK (mean ± SD)	242.4 ± 15.5 (n=26)	244.1 ± 13.2[b] (n=25)	218.8 ± 11.4 (n=6)	224[c] (n=1)
Residency Placement [n(%)]				
Matched in top choice	55 (88.7%)	55 (91.7%)	5 (25%)[d]	3 (75%)
Matched in 2nd choice	6 (9.7%)	5 (8.3%)	11 (55%)[e]	0
Unmatched	1 (1.6%)	0	4 (20%)	1 (25%)

[a] $P < .05$ between Agreement on Promotion and Model 2 down-classified groups.

[b] $P < .001$ between Agreement on Promotion and Model 2 down-classified groups.

[c] SD is missing because only one student took the USMLE in this group.

[d] $P < .001$ between Agreement on Promotion, Model 2 down-classified and Model 1 down-classified groups.

[e] $P < .001$ between residency placement categories within Model 2 down-classified column.

Abbreviations: USMLE: United States Medical Licensing Examination; CK: Clinical Knowledge.

TABLE XIII

COMPARISON BETWEEN MODELS 1 AND 3: USMLE SCORES AND RESIDENCY PLACEMENT AMONG DIFFERENT AGREEMENT/DISAGREEMENT GROUPS

			Disagreement	
	Total agreement (n=75)	Agreement on Promotion (n=71)	Model 3 down-classified cases (n=7)	Model 1 down-classified cases (n=4)
USMLE Step 1 (mean ± SD)	229.9 ± 16 (n=39)	229.9 ± 16 (n=39)	173.0[a] (n=1)	221[a] (n=1)
USMLE Step 2 CK (mean ± SD)	240.2 ± 15.6 (n=30)	240.2 ± 15.6 (n=30)	210.0[a] (n=1)	212.5 ± 16.3 (n=2)
Residency Placement [n(%)]				
Matched in top choice	59 (78.7%)	57 (80.3%)[b]	3 (42.9%)	1 (25%)
Matched in 2nd choice	14 (18.7%)	13 (18.3%)	2 (28.6%)	1 (25%)
Unmatched	2 (2.7%)	1 (1.4%)	2 (28.6%)	2 (50%)

[a] SD is missing because only one student took the USMLE in this group.
[b] *P* < .01 between Agreement on Promotion, Model 3 down-classified and Model 1 down-classified groups.
Abbreviations: USMLE: United States Medical Licensing Examination; CK: Clinical Knowledge.

3. Comparison between models 2 and 3

Comparison of USMLE scores across agreement/disagreement groups using ANOVA showed significant difference for USMLE Step 1 and Step 2 CK ($F_{(2,38)}$ = 7.418, *P* = .002 for USMLE Step 1 scores and $F_{(2,30)}$ = 11.180, *P* < .0001 for USMLE Step 2 CK scores). Pair-wise comparison using the independent-samples *t* test (post hoc analyses were not possible with the small sample size noted within some groups) showed that, on average, students who were down-classified in model 2 had non significantly lower Step 1 USMLE scores by comparison to those who were granted promotion in the 2 models (equal-variance $t_{(38)} = 1.39$, $P = .17$), and significantly lower Step 2 CK USMLE scores (equal-variance $t_{(30)} = 4.25$, $P < .0001$). Overall, there was a significant difference in residency placement percentages across agreement/disagreement groups ($\chi^2_{(4)} = 22.691$, $P < .0001$). Post hoc analyses showed that students who were down-classified in model 2 were less likely to match in first choice than to match in second choice or not to match ($\chi^2_{(2)} = 26.86$, $P < .0001$), and they were less likely to match in first choice by comparison to students who were promoted in both models and those who were down-classified in model 1 ($\chi^2_{(2)} = 20.79$, $P < .0001$). By comparison, students who were down-classified in model 3 had equal chances to fall into any of the residency categories (Table XIV).

TABLE XIV

COMPARISON BETWEEN MODELS 2 AND 3: USMLE SCORES AND RESIDENCY PLACEMENT AMONG DIFFERENT AGREEMENT/DISAGREEMENT GROUPS

	Total agreement (n=63)	Agreement on Promotion (n=59)	Disagreement	
			Model 3 down-classified cases (n=6)	Model 2 down-classified cases (n=17)
USMLE Step 1 (mean ± SD)	231.1 ± 15.9 (n=34)	231.1 ± 15.9 (n=34)	173[a] (n=1)	221.5 ± 13.3 (n=6)
USMLE Step 2 CK (mean ± SD)	243.3 ± 13.6 (n=26)	243.3 ± 13.6[b] (n=26)	210[a] (n=1)	217.3 ± 13.2 (n=6)
Residency Placement [n(%)]				
Matched in top choice	53 (84.1%)	53 (89.8%)	5 (83.3%)	5 (29.4%)[c]
Matched in 2nd choice	8 (12.7%)	5 (8.5%)	0	9 (52.9%)[d]
Unmatched	2 (3.2%)	1 (1.7%)	1 (16.7%)	3 (17.6%)

[a] SD is missing because only one student took the USMLE in this group.

[b] *P* < .001 between Agreement on Promotion and Model 2 down-classified groups.

[c] *P* < .001 between Agreement on Promotion, Model 3 down-classified and Model 2 down-classified groups.

[d] *P* < .001 between residency placement categories within Model 2 down-classified column.

Abbreviations: USMLE: United States Medical Licensing Examination; CK: Clinical Knowledge.

IV. DISCUSSION

The major findings of this study are related to validity evidence associated with the use of composite scores and decision-making processes in the context of competency-based undergraduate medical education. These findings can be framed around the 3 sources of validity evidence presented below.

A. Internal Structure Validity Evidence

1. Reliability of composite scores

Since many constructs are involved in the accomplishment of professional tasks, assessment is expected to reflect, not only individual construct-related performance, but ideally, the global performance resulting from the interfacing of constructs as well. Irrespective of how assessment information is used, a minimum set of quality standards should be respected and secured. Frequently in medial education, people (administrators, students, educators, etc.) rely on this information to take a series of actions, from providing feedback to making significant decisions about students' selection or academic progress. Since assessment information about a medical student is multi-layered and complex, the practice has been to group this information and present it in a meaningful way to serve the intended purposes. Combining assessment scores into composites has been shown to be a successful method of aggregating information about a student's performance (41). However, the educational system in place determines the framework that drives data assembly. This understanding challenged the procedure of combining assessment information in competency-based medical education, because the whole focus shifted from achieving an objective to demonstrating competence in a domain and inspiring trust in a professional task. In this study we demonstrated that the reliability of composite scores, formulated according to either the traditional medical education model or the competency-based model, is adequate for the scope of use of these scores, which adds an internal structure validity evidence supporting inferences (and decisions) made from these scores (42).

2. **Inter-rater reliability and rater rationales**

Despite the different rationales presented by the 8 educators, they were all in very good agreement about which students should be promoted and who would require further discussion, as noted in the high intraclass correlation coefficient (40). This could be explained by the fact that an underperforming student is underachieving in more than one domain, which makes identification of the student a high possibility irrespective of the rationale or decision model.

Analyses of the qualitative data yielded a surprising observation that very few members approached the selection task from a competency-based perspective. Knowing that the school had recently involved most educators (including members of this committee) in the formulation of outcome competencies for the recently introduced competency framework, we expected to find more competency-based rationales. This finding confirms that our educators haven't yet built a clear vision about the model. However, these results, in addition to the quantitative ones, should be considered in planning the change from the traditional medical education model to the competency-based model, as outlined in the 8-step process for implementing a change by Kotter (43). This model brought an additional perspective to decision-making: a deliberation process that involves consensus, values narrative assessment, and de-emphasizes composite score computation. All members considered narrative assessment at some point during their analysis. This contributed to decision-making in 2 ways: first, by re-considering the assessment of Professionalism, and second, by raising concerns over the rater bias introduced in CPE scores. Some members selected students based on negative comments about Professionalism, which did not coincide with the numerical rating attributed to this domain. Other members had similar observations about the rating of items on the CPE form, where contradictions were noted between included comments and numerical ratings for some items. These observations indicate that some areas in our assessment system need to be reconsidered and improved, in particular through re-thinking the assessment of Professionalism, introducing targeted rater training, and encouraging the practice of failing weak students.

This qualitative study, through the analysis of consensus building, added another evidence supporting decisions made by this model. Moreover, it improved our understanding of faculty concerns and helped identify areas for improvement in our assessment system.

B. Relations to Other Variables

The measure of a specific construct is expected to correlate well with another measure of the same construct and less well with a measure of a different construct. Demonstrating discriminant correlations provides additional validity evidence supporting inferences made from assessment scores. In our study, knowledge tests (MCQs) significantly correlated with each others across clerkships, and associated less well with other tests. However, correlation was very variable and clerkship-specific for the rest of the tests. This indicates that performance tests, including OSCE and CPE, are probably not measuring homogeneously the same constructs across clerkships (44). For example, a Pediatric OSCE may focus more significantly on communication skills than on other skills by comparison to an Internal Medicine OSCE that may focus on overall skills. In this case, the Pediatric OSCE score would correlate less well with the Pediatric CPE score, while the opposite would occur for IM. This finding reflects the instability of construct representativeness, another area for re-consideration and possible improvement in our assessment system.

USMLE scores correlated significantly with MCQ examination scores and very variably with performance tests, depending on the clerkship. As expected, USMLE Step 2 CK scores correlated more significantly than Step 1 scores with most clerkship composite scores, particularly because both scores reflect more clinical than basic science abilities. Concerning competency-based composite scores, they correlated to varying extent among each others and with USMLE scores. More specifically, 'Physician as Advocate' and 'Physician as Care Giver' composite scores correlated with all other composites, mainly because these competency domains require the integration of different abilities, such as knowledge, communication skills, professional behavior, in addition to others. On the other hand, the 'Physician as Scientist' composite score correlated less well with the 'Physician as Communicator' and 'Physician as Professional' composites, a finding that is expected given the different constructs that these roles involve. USMLE scores, which are measures of knowledge mostly, correlated significantly with the 'Physician as Scientist' and 'Physician as Care Giver' composite scores. Our results showed that discriminant correlations are more meaningful and interpretable in this decision model.

C. Consequential Validity Evidence

This source of validity evidence is probably the most critical in medical education because the consequences of assessment affect the student, the educators and administrators, but most importantly they affect the patient. Re-thinking the medical profession revealed a mismatch between the education of future physicians and the requirements of the profession (45). In particular, graduates were frequently unprepared to exercise without supervision. CBME and Entrustable Professional Activities frameworks were brought forward to re-align education and professional practice, and to reframe decision-making processes around professional expectations (12). One of the most challenging tasks in medical education is to make defensible decisions about students' advancement, more essentially decisions for borderline cases (46-50). Such decisions are of particular importance because underachieving students usually maintain a borderline performance throughout their studies (51). Additionally, this performance frequently fluctuates between the unsatisfactory and satisfactory levels, which increases the risk of false positive and false negative decisions in this performance spectrum. There is a common belief that interventions to remediate deficiencies result in significant performance improvement (51-55), hence the raised concern over the failure to identify struggling trainees. Pass/Fail decisions or performance classification frequently rely on cutoff scores that need to be well thought of and optimized to avoid (or minimize) incorrect decisions. In our study, we examined classification accuracy in three models, as a source of consequential validity evidence, by testing the quality of classification associated with each model and external criteria of performance. These models involved the traditional use of composite scores, the use of re-formulated composite scores guided by the competency framework, and deliberations among education experts, also guided by the competency framework. The use of proper standard setting procedures for each assessment and assessment component whenever applicable (such as each OSCE station), and the application of score scaling before combining scores, represent one source of evidence supporting the pass/fail decision validity. Furthermore, our results showed that the 3 models identified underperforming students, despite a differential re-classification of some students. Decision model 1 identified 11.6% of the students as either promotable but requiring academic monitoring or requiring a remedial work before promotion is considered. The rest of the students (88.4%) were granted an unconditional promotion. The latter group demonstrated a significantly better performance on USMLE Step 2 CK (borderline significance) compared to borderline students and a higher chance of matching in top residency choice. On the other hand, model 2 significantly identified more students in the 'Borderline' decision category by comparison to the other models, and these students had significantly lower USMLE Step 1 and Step 2 CK scores and lower chances to match in first residency choice by comparison to promoted students. Similarly, model 3 identified more borderline students than model 1, who were less likely to match in first residency choice. However, students in this group did not complete any of the USMLE, which limited further studies. Therefore, the 3 models were able to identify borderline students who were under-achieving by comparison to promoted students, although the strength of the associations with the external criteria was more significant in model 2 as compared to models 1 and 3.

Identifying struggling students has implications not only for students' academic advancement, but also for the faculty and the institution. Remediation strategies have to be tailored to the identified deficiencies and they require human resources and time commitment. Therefore, decisions over borderline students should be well informed and guided by validity studies. To further examine the differential re-classification of the 3 models, we explored outcome characteristics of groups that were down-classified by each model (disagreement analysis). Our aim was to investigate whether down-classified students were underperformers, which would privilege the use of one model over the other. Our findings showed that students who were down-classified in model 2 had significantly lower performance on USMLE and lower chances to match in first residency choices compared to those whose promotion status resulted from common agreement between models. These results were not as striking in down-classified groups of the other 2 models, where some statistics were limited by the small sample size. These findings suggest that model 2 was more likely to identify real underperformers by comparison to the other 2 models.

Prior studies have shown that struggling students may represent between 0% and 15% of a cohort (55, 56). In our study, combining the 2 middle decision categories (Borderline Performance and Remedial), we found that model 1 identified 10 students (11.6%) as struggling learners, model 2 identified 20 students (23.3%), and model 3 identified 12 students (13.9%). Despite these variable percentages, the three models classified 58 students (67.4%) similarly. The remaining 28 cases were either down-classified or up-classified, depending on the model. One possible reason behind the lower percentage of borderline students in model 1 compared to other models is that faculty are frequently reluctant to fail students based on clinical evaluation of performance for many reasons that are beyond the scope of this study (57, 58). Given that model 1 decision policies mandate a passing CPE score and a compensatory approach involving other assessment tests (MCQ and OSCE), struggling students on these latter tests were less likely to be identified in this model, and more likely in model 2. Furthermore, our definition of borderline performance was inclusive of students who were only on the passing facet of a borderline performance and without any significant failure, which increased the chances of identifying more students in the three models.

Finally, our results showed that decisions ensuing from the three models are supported by consequential validity evidence, and that model 2, using competency-guided composite scores, provided a better classification accuracy, especially in the borderline spectrum of performance.

D. Strengths and Limitations of the Study

Strengths of this study are substantial. Presenting multiple sources of validity evidence associated with the use of composite scores in decision-making reinforces conclusions and facilitates their acceptability. Additionally, the involvement of education leaders in the formulation of decisions improved their understanding of the educational system and created a further commitment towards education, as expressed in heir follow-up comments. The process used to address the challenge helped in many ways in the identification of gaps in our assessment system. First, the exercise of re-computing composite scores based on the competency framework was based on a mapping of CanMEDS Roles to assessment tests. Through this mapping, we realized that the 'Physician as Advocate' role for example was not assessed sufficiently and in depth by comparison to other roles. Second, the inclusion of a qualitative study enriched our results by revealing the perspectives of education leaders and their concerns about assessment practices at the school. Another strength is the multi-faceted analysis of results. Exploring how decision models disagreed about cases and analyzing the effect of this disagreement on student re-classification strengthened our conclusions about a preferential use of models. Finally, results of this study will be used to implement improvements, initiate educational change, and widen the research agenda of faculty.

This study has many limitations. First, its monocentric model (single institution experience) limits the generalizability of our results. However, the nature of the challenge being common to many educational institutions makes our study plausible for replication in different contexts. Moreover, the process followed to address the challenge can be generalizable to other places as well. Second, the small number of available USMLE scores limited some statistical analyses, which could have affected our results. Another limitation is related to the fact that most residency placements were intra-mural, where the selection system relies on reported traditional composite scores. This could have favored the first model.

E. Implications for Research

This study presented substantial evidence supporting the use of composite scores in decision-making about trainees' advancement. Furthermore, it presented a demonstration about the formulation of composite scores as measures of competencies. However, some questions remained unanswered and others emerged in light of this study. How would our results change should all students have USMLE scores? How would the identification of a wider pool of trainees as underperformers affect the system and how would targeted interventions affect future performance? Given that the third model was more inclusive of assessment information but short of a common selection process, what would be the effect of using a blend of models, by starting with a first selection using model 2 and then supporting decisions using narrative information?

F. Conclusions

Our study demonstrated that the use of composite scores is associated with defensible decisions about student advancement, irrespective of the medical education model in which they were formulated. However, decision models differ with their ability to address the challenge of identifying struggling students. Although the advancement of CBME had implications over assessment, formulating composite scores as measures of competencies is feasible and yields better classification decisions. A competency-based decision model could better identify students with borderline performance by internal and external criteria, which would facilitate interventions. On the other hand, deliberations among education experts may improve decisions based on traditional models but would ultimately serve as a complementary process supporting an existing decision model.

CITED LITERATURE

1. Kane, M., and Case, S.M.: The reliability and validity of weighted composite scores. *Applied Measurement in Education,* 17(3): 221-240*,* 2004.

2. Carman, K.: Improving quality information in a consumer-driven era: Showing differences is crucial to informed consumer choice. Presentation delivered at the 10th National CAHPS User Group Meeting. 2006 Mar 31. Available at http://archive.ahrq.gov/cahps/news-and-events/events/UGM10/DAY2_cd_1_Carman.pdf.

3. Crone, L.J., Lang, M.H., Franklin, B.J., and Halbrook, A.M.: Composite versus component scores: Consistency of school effectiveness classification. *Applied Measurement in Education,* 7(4): 303-321, 1994.

4. Corcoran, J., Downing, S.M., Tekian, A., and DaRosa, D.A.: Composite score validity in clerkship grading. *Academic Medicine,* 84 (10 suppl): S120-S123, 2009.

5. Park, Y.S., Lineberry, M., Hyderi, A., Bordage, G., Xing, K., and Yudkowsky, R.: Differential weighting for subcomponent measures of Integrated Clinical Encounter scores based on the USMLE Step 2 CS examination: Effects on composite score reliability and pass-fail decisions. *Academic Medicine,* 91: S24-S30, 2016.

6. Moonen-van Loon, J.M.W., Overeem, K., Donkers, H.H.L.M., van der Vleuten, C.P.M., and Driessen, E.W.: Composite reliability of a workplace-based assessment toolbox for postgraduate medical education. *Adv in Health Sci Educ,* 2013, DOI 10.1007/s10459-013-9450-z.

7. Kreiter, C.D., and Bergus, G.R.: A study of two clinical performance scores: Assessing the psychometric characteristics of a combined score derived from clinical evaluation forms and OSCEs. *Medical Education Online,* 12: 10, 2007.

8. American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (U.S.). Standards for educational and psychological testing. Washington, DC: AERA, 2014.

9. Frank, J.R., Snell, L.S., ten Cate O., Holmboe, E.S., Carraccio, C., Swing, S.R., Harris, P., Glasgow, N.J., Campbell, C., Dath, D., Harden, R.M., Lobst, W., Long, D.M., Mungroo, R., Richardson, D.L., Sherbino, J., Silver, I., Taber, S., Talbot, M., and Harris, K.A.: Competency-based medical education: theory to practice. *Medical Teacher,* 32(8): 638-645, 2010.

10. Holmboe, E.S., Sherbino, J., Long, D.M., Swing, S.R., and Frank, J.R.: The role of assessment in competency-based medical education. *Medical Teacher,* 32(8): 676-682, 2010.

11. Harris, P., Snell, L., Talbot, M., and Harden, R.M.: Competency-based medical education: implications for undergraduate programs. *Medical Teacher,* 32(8): 646-650, 2010.

12. ten Cate, O.: Entrustability of professional activities and competency-based training. *Medical Education*, 39(12): 1176-1177, 2005.

13. Kelly, T.: Interpretation of Educational Measurements. New York, World Book Company, 1927.

14. Gulliksen, H.: Theory of mental tests. New York, Wiley, 1950.

15. Jarjoura, D., and Brennan, R.: A variance components model for measurement procedures associated with tables of specifications. *Applied Psychological Measurement,* 6: 161-171, 1982.

16. Scriven, M.: Validity in personnel evaluation. *Journal of Personnel Evaluation in Education,* 1: 9-23, 1987.

17. Hambleton, R., and Slater, S.: Reliability of credentialing examinations and the impact of scoring models and standard-setting policies. *Applied Measurement in Education,* 10: 19-38, 1997.

18. Feldt, L.: Can validity rise when reliability declines? *Applied Measurement in Education*, 10: 377-387, 1997.

19. Brennan, R.: Some problems, pitfalls, and paradoxes in educational measurement. *Educational measurement: Issues and Practice*, 20(4): 6-18, 2001.

20. Downing, S.M.: Statistics of testing. In: Assessment in Health Professions Education, eds. S.M. Downing and R. Yudkowsky, pp. 93-117. Routledge, NY, 2009.

21. Kolen, M.J., and Brennan, R.L.: Score Scales. In: Test equating, Scaling and Linking. Methods and Practices, eds. S.E. Fienberg and W.J. van der Linden, pp. 371-424. Springer, New York, 2014.

22. Rudner, LM.: Informed test component weighting. *Educational Measurement: Issues and Practice,* 20(1): 16–19, 2001.

23. Wang, M.W., and Stanley, J.C.: Differential weighting: A review of methods and empirical studies. *Review of Educational Research*, 4: 663–704, 1970.

24. Stagnaro-Green, A., Deng, W., Downing, S.M., and Crosson J.: Theoretical model evaluating the impact of weighted percent versus standard scores in determining third year clerkship grades. Paper presented at the Annual Meeting of the Association of American Medical Colleges, RIME, Boston, MA, 2004.

25. Yudkowsky, R., Tumuluru, S., Casey, P., Herlich, N., and Ledonne, C.: A patient safety approach to setting pass/fail standards for basic procedural skills checklists. *Simulation in Healthcare,* 9(5): 277-282, 2014.

26. Reinert, A.: Assessment in medical education: A primer on methodology [Online]. College of Physicians and Surgeons, Columbia University. Accessed on December 7th, 2016. Available from: http://ps.columbia.edu/education/edu-news-archive/article-anna-reinert-md-class-2013.

27. Park, Y.S., Hodges, B., and Tekian, A.: Evaluating the paradigm shift from time-based toward competency-based medical education: implications for curriculum and assessment. In: Assessing Competence in Professional Performance Across Disciplines and Professions, eds. P. Wimmers and M. Mentkowski, pp. 411-425. Springer International Publishing Switzerland, 2016.

28. Cureton, E.E.: Validity. In: Educational Measurement, ed. E.F. Lindquist, pp. 621-694. Washington, DC, American Council on Education, 1951.

29. Ebel, R.: Must all tests be valid? *American Psychologist,* 16: 640-647, 1961.

30. Cronbach, L.J.: Test validation. In: Educational Measurement, 2nd Ed., ed. R.L. Thorndike, pp. 443-507. Washington, DC, American Council on Education, 1971.

31. Messick, S.: Validity of psychological assessment: Validation of inferences from persons’ responses and performances as scientific inquiry into score meaning. *American Psychologist,* 50 (9): 741–9, 1995.

32. Kane, M.: Validation. In: Educational Measurement, 4th Ed., ed. R.L. Brennan, pp. 17-64. Westport, CT, American Council on Education/Praeger, 2006.

33. Kane, M.T.: Validating the interpretations and uses of test scores. *Journal of Educational Measurement,* 50: 1-73, 2013.

34. Eva, K.W., and Reiter, H.I.: Where judgment fails: Pitfalls in the selection process for medical personnel. *Advances in Health Sciences Education,* 9: 161-174, 2004.

35. Leveille, D.E.: Accountability in Higher Education: A Public Agenda for Trust and Cultural Change. Berkeley: Center for Studies in Higher Education, 2006.

36. Swan, G.: Tools for Data-driven Decision Making in Teacher Education: Designing a Portal to Conduct Field Observation Inquiry. *Journal of Computing in Teacher Education*, 25 (3): 107-113, 2009.

37. Andolsek, K., Padmore, J., Hauer, K.E., and Holmboe, E.: Clinical Competency Committees. A guidebook for programs. The Accreditation Council for Graduate Medical Education, 2015.

38. Harris, I.: What does "The Discovery of Grounded Theory" have to say to medical education? *Advances in Health Sciences Education,* 8: 49-61, 2003.

39. Strauss, A., and Corbin, J. Basics of qualitative research: Grounded theory procedures and techniques. Newbury Park, CA, Sage, 1990.

40. Landis, J.R., and Koch, G.G.: The measurement of observer agreement for categorical data. *Biometrics,* 33: 159–74, 1977.

41. Bobko, P., Roth, P.L., and Buster, M.A.: The usefulness of unit weights in creating composite scores. A literature review, application to content validity, and meta-analysis. *Organizational Research Methods,* 10: 689-709, 2007.

42. Downing, S.M.: Reliability: On the reproducibility of assessment data. *Medical Education,* 2004, 38: 1006-1012.

43. Kotter, J.P.: Leading Change. First Ed. Boston, MA, Harvard Business Press, 1996.

44. Kies, S.M., Roth, V., and Rowland, M.: Association of third-year medical students' first clerkship with overall clerkship performance and examination scores. *Journal of the American Medical Association,* 304(11): 1220-1226, 2010.

45. Cooke, M., Irby, D.M., Sullivan, W., and Ludmerer, K.M.: American medical education 100 years after the Flexner report. *New England Journal of Medicine,* 355: 1339-1344, 2006.

46. Shulruf, B., Turner, R., Poole, P., and Wilkinson, T.: The objective borderline method (OBM): A probability-based model for setting up an objective

pass/fail cut-off score for borderline grades in medical education programmes. *Advances in Health Sciences Education*, 18 (2), 231-244, 2013.

47. Shulruf, B., Jones, P., and Turner, R.: Using student ability and item difficulty for making defensible pass/fail decisions for borderline grades. *Higher Education Studies,* 5: 106-118, 2015.

48. Schoonheim-Klein, M., Muijtjens, A., Habets, L., Manogue, M., Van der Vleuten, C., and Van der Velden, U.: Who will pass the dental osce? Comparison of the angoff and the borderline regression standard setting methods. *European Journal of Dental Education*, 13 (3), 162-171, 2009.

49. Subkoviak, M.J.: A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement,* 25: 47-55, 1988.

50. Douglas, K.M., and Mislevy, R.T.: Estimating classification accuracy for complex decision rules based on multiple scores. *Journal of Educational and Behavioral Statistics,* 35: 280-306, 2010.

51. Pell, G., Fuller, R., Homer, M., and Roberts, T.: Is short-term remediation after osce failure sustained? A retrospective analysis of the longitudinal attainment of underperforming students in osce assessments. *Medical Teacher,* 34 (2), 146-150, 2012.

52. Faustinella, F., Orlando, P.R., Coletti, L.A., Juneja, H.S., and Perkowski, L.C.: Remediation strategies and students' clinical performance. *Medical Teacher*, 226: 664 – 665, 2004.

53. Wiese, J.G.: A curriculum to observe underachieving students and give assisted remediation (COUGAR). *Journal of General Internal Medicine*, 15(suppl. 1): 224, 2000.

54. Segal, S.S., Giordani, B., Gillum, L.H., and Johnson, N.: The academic support program at the University of Michigan School of Medicine. *Academic Medicine*, 74: 383–385, 1999.

55. Frellsen, S.L., Baker, E.A., Papp, K.K., and Durning, S.J.: Medical school policies regarding struggling medical students during the internal medicine clerkships: results of a National survey. *Academic Medicine,* 83: 876-881, 2008.

56. Yates, J., and James, D.: Predicting the "strugglers": A case-control study of students at Nottingham University Medical School. *British Medical Journal*, 332:1009 –1013, 2006.

57. Guerrasio, J., Furfari, K.A., Rosenthal, L.D., Nogar, C.L., Wrag, K.W., and Aagaard, E.M.: Failure to fail: The institutional perspective. *Medical Teacher,* 36:799-803, 2014.

58. Yepes-Rios, M., Dudek, N., Duboyce, R., Curtis, J., Allard, R.J., and Varpio, L.: The failure to fail underperforming trainees in health professions education: A BEME systematic review: BEME guide No. 42. *Medical Teacher,* 38:1092-1099, 2016.

APPENDICES

Appendix A

Program Outcome Competencies

Clinical Competencies:

During their clinical rotations, students will continuously acquire knowledge, skills and attitudes that will shape-up their personality as future caregivers, and prepare them for the practice of medicine in the 21st century. Competency-based education is currently used worldwide and will help us frame our clinical education and assessment. At LAU Gilbert and Rose-Marie Chagoury School of Medicine (SOM) during Medicine years I and II, the educational program embraced four competency related themes: Basic and Clinical Sciences, Clinical Competence, Professional and Behavioral, and Social Medicine and Public Health. These themes will continue to be woven into the teaching and learning of Medicine in years III and IV. By the end of medical school students are expected to achieve a level of proficiency that will not only prepare them for graduate medical education, but also build the foundation for lifelong learning.

Building upon these themes, the SOM is committed to preparing graduates who demonstrate achievement in the following basic competencies in the practice of medicine, derived from the CanMEDS Roles:

PHYSICIAN AS SCIENTIST
PHYSICIAN AS COMMUNICATOR
PHYSICIAN AS CARE GIVER
PHYSICIAN AS ADVOCATE
PHYSICIAN AS PROFESSIONAL

PHYSICIAN AS SCIENTIST

At the end of Medicine year III the student will be expected to:

1. **Possess a solid foundation of scientific and medical knowledge and apply this knowledge to the care of patients and populations**

The following are behaviors or practices that the student adopts and applies to demonstrate competence:

- Describes the normal structure and function of the human body

- Explains pathologic and pathophysiologic processes leading to alterations in normal structure and function of the human body for major causes of illness
- Describes population based factors that affect disease prevalence, course and treatment
- Describes patterns of diseases at different levels of acuity: emergency, acute and chronic
- Discusses social determinants of health and illness
- Explains principles of pharmacology and major categories of drugs, their actions, interactions, contraindications and clinical uses.
- Explains the principles underlying normal behavior and mental diseases
- Solves basic clinical problems using knowledge of mechanisms of disease

Measures of assessment include:

- Multiple Choice Questions (MCQs) exams
- Clinical Performance Evaluation (CPE)
- Objective Structured Clinical Examination (OSCE)
- Presentations

2. **Continue to seek, access, analyze and apply knowledge to a changing environment**

Behaviors that demonstrate competence:

- Contributes new knowledge to health care team
- Utilizes evidence-based medicine for patient management
- Seeks appropriate resources for improving quality of care
- Critically analyzes literature

Measures of assessment include:

- CPE
- Written and oral presentations
- Contributions on rounds
- Presentations and projects
- Mini-Clinical evaluation exercise (mini-CEX)

PHYSICIAN AS COMMUNICATOR

At the end of Medicine year III the student will be expected to:

1. **Communicate effectively with patients, their families, colleagues and the health care team.**

Behaviors that demonstrate competence:

- Develops effective patient-physician relationship, showing respect and empathy
- Listens attentively and elicits appropriate data in history taking
- Elicits pertinent social, cultural and economic determinants of health
- Effectively exchanges data both verbally and in writing with members of the health care team
- Demonstrates ability to use appropriate communication skills in discussing diagnosis and disease management with patients
- Is an active valued member of the healthcare team
- Presents patient information clearly, accurately and in a timely fashion
- Demonstrates ability to use all communication skills both verbally and in writing
- Involves patient and family in developing healthcare plan
- Educates patient on illness and care plan
- Practices coordinated care among members of the healthcare team

- Cooperates with patients and healthcare professionals from diverse cultural backgrounds

Measures of assessment include:

- CPE, mostly history taking, presentation skills and writing skills
- Multisource 360° assessment by patient and other health care providers
- Projects and reports
- Critical incident Report
- OSCE
- Mini CEX

PHYSICIAN AS CARE GIVER

At the end of Medicine year III the student will be expected to:

1. **Practice evidence based medicine in the care of patients**

Behaviors that demonstrate competence

- Recognizes life-threatening situations and delivers basic emergency care for such patients within or outside healthcare facilities
- Performs both complete and focused physical examination
- Synthesizes data to formulate a differential diagnosis
- Describes the appropriate use of laboratory and radiologic techniques in identifying health problems
- Orders (Mock) appropriate diagnostic tests in correct sequence
- Describes the appropriate use of laboratory and radiologic techniques in identifying health problems

- Writes patient care and management plan based on scientific principles, evidence based approach and guidelines from scientific societies in a compassionate manner
- Discusses both pharmacological and non- pharmacological management plans
- Re-evaluates patient status and management plan
- Meets all technical skills requirements
- Writes discharge summary and plan
- Applies prescription skills to discharge planning and medication reconciliation

Measures of assessment include:

- CPE
- Mini-CEX
- OSCE
- MCQs exam
- Log book

2. **Utilize the full spectrum of health care delivery: acute, chronic, preventive, rehabilitative, public health and social services to optimize individual and population based care**

Behaviors that demonstrate competence:

- Seeks care solutions through various forms of health services
- Describes various levels of care for each patient
- Describes healthcare delivery systems in Lebanon and other countries
- Integrates knowledge of social, cultural and behavioral factors as well as preventive measures and cost effective analysis to advance patient well-being

Measures of Assessment include:

- CPE
- OSCE
- Mini-CEX
- MCQs

PHYSICIAN AS ADVOCATE

At the end of Medicine year III the student will be expected to:

1. **Advocate for improved health care: access, health outcomes, health promotion and disease prevention, community services**

Behaviors that demonstrate competence:

- Identifies social determinants of health
- Seeks solutions to barriers for access to health care
- Recognizes the impact of money and industry on the practice of medicine
- Discusses community-driven plans for health promotion and disease prevention
- Discusses the clinical encounter from a cross-disciplinary perspective

Measures of assessment include:

- Continuity experience write-up
- Community experience in primary care clerkship
- Projects
- CPE

PHYSICIAN AS PROFESSIONAL

At the end of Medicine year III the student will be expected to:

1. **Consistently practice and model ethical and professional behavior**

Behaviors that demonstrate competence:

- Acts in the patient's best interest
- Demonstrates reliability and responsibility, and performs duties always in a timely and dependable manner
- Follows rules of health care facility and code of conduct
- Demonstrates respect and application of policies that govern clinical practice in the country
- Respects patients, family, colleagues, other health care providers and employees
- Respects rights of patient and family
- Applies ethical principles to decision making in patient care
- Educates patient and family on informed consent
- Describes Institutional Review Board (IRB) process for research

Measures of assessment include:

- CPE
- Multisource 360° assessment
- OSCE

2. **Reflect on practice and ways to improve self, patient safety and quality**

Behaviors that demonstrate competence:

- Analyzes personal experience, acknowledge gaps and work on ways to improve them
- Discusses medical errors, quality improvement opportunities and participates in solutions to reduce them

- Writes critical incident reports
- Demonstrates responsibility for continuous learning and personal growth and development
- Identifies areas of weakness and educational needs, and develops an improvement plan using appropriate learning resources

Measures of assessment include:

- Critical incident report
- CPE

Appendix B

Clinical Performance Evaluation Form

LAU School of Medicine
Clerkship Clinical Performance Evaluation Form (Global Rating)

Please complete the form using black or blue ink. Use the cross sign "x" when applicable. Mark "NA" for non-applicable if a judgment cannot be made clearly or if the area of competency was not assessed.

Student Name: Date of Evaluation:

Clerkship Site: Evaluator:

Months spent in clerkship:

Please indicate the extent of contact you had with the medical student: hours.

For each of the areas or competencies listed below, please check the appropriate level of ability.

<table>
<tr><td colspan="4"><b>1. FUND OF KNOWLEDGE – Demonstrates knowledge of topics, application of basic science and clinical knowledge to patient care and ability to access high quality data.</b> <b>NA □</b>
Please indicate the basis on which you are evaluating the student. Check all that apply.
Rounds □ Conferences □ Presentations □ Precepting on clinical cases □ Other _______</td></tr>
<tr><td>Below Expectations
□</td><td>Needs Significant Improvement
□</td><td>Meets Expectations
Passably □ Sufficiently □</td><td>Exceeds Expectations
Excellently □ Outstandingly □</td></tr>
<tr><td><b>Often</b> significant gaps in knowledge, limited understanding of pathophysiology, diagnosis or management, unable to correlate basic and clinical sciences to patient care, little evidence of seeking of reliable information</td><td><b>Inconsistently</b> demonstrates substantial foundation in applied knowledge with frequent fluctuation between "Below Expectations" and "Meets Expectations" performance, most importantly needs significant improvement before moving to a</td><td><b>Usually</b> solid fund of basic and clinical science knowledge related to care of patient, usually reads about patient's condition, usually accesses information from high quality resources</td><td><b>Consistently</b> applies basic and clinical science knowledge to care of patient, provides evidence based data, seeks information from high quality resources, major contributor to knowledge of health care team, grasps medical controversies</td></tr>
</table>

	higher academic level		

2. HISTORY TAKING – Demonstrates ability to take a complete and/or focused history. NA □

Please indicate the basis on which you are evaluating the student.

Observed history taking □ Student presentations □ Number of times _______ Number of times _______

Below Expectations	Needs Significant Improvement	Meets Expectations	Exceeds Expectations
□	□	Passably □ Sufficiently □	Excellently □ Outstandingly □
Often disorganized, misses information, inappropriate questions, poorly focused, out of sequence, inaccurate data	**Inconsistently** organized and complete with frequent fluctuation between "Below Expectations" and "Meets Expectations" performance, most importantly needs significant improvement before moving to a higher academic level	**Usually** organized, logical and complete, focused with accurate present illness and past medical history, identifies major relevant components and new problems, elicits pertinent (ROS) Review of Systems both positive and negative, elicits pertinent psychosocial and personal history	**Consistently** organized , logical, complete, focused, accurate, with past medical history, + and – ROS, psychosocial and personal history, reflects understanding of disease course and of the patient's circumstances

3. PHYSICAL EXAMINATION – Demonstrates ability to perform a complete/ or focused physical examination of the patient. NA □

Please indicate the basis on which you are evaluating the student. Check all that apply.

Observed complete physical examination □ Number of times___ Observed focused physical examination □ Number of times___

Precepting on clinical cases □ Other ……..

Below Expectations	Needs Significant Improvement	Meets Expectations	Exceeds Expectations
□	□	Passably □ Sufficiently □	Excellently □ Outstandingly □
Often disorganized, incomplete or unreliable exam, missing major findings, inability to correlate physical exam with history, uncertainty of normal and abnormal findings, insensitive to patient comfort	**Inconsistently** organized and complete exam with frequent fluctuation between "Below Expectations" and "Meets Expectations" performance, most importantly needs significant improvement before moving to a higher academic level	**Usually** organized, systematic exam, complete or focused as correlated with history, identifies major findings, recognizes normal and abnormal, sensitive to patient	**Consistently** and efficiently performs organized physical examination either complete or focused as related to history, sensitive to patient, distinguishes normal form abnormal, elicits subtle findings

4. COMMUNICATION SKILLS – Demonstrates ability to communicate effectively with patients, families and colleagues.
NA □

Please indicate the basis on which you are evaluating the student. Check all that apply.

Rounds □ Conferences □ Presentations □ Precepting on clinical cases □ Other ……..

Below Expectations □	Needs Significant Improvement □	Meets Expectations Passably □ Sufficiently □	Exceeds Expectations Excellently □ Outstandingly □
Often unable to communicate with patient or health care team, inadequate listening skills, inadequate sensitivity to patient, judgmental, overuse of medical jargon understandable to patient, failure to develop rapport	**Inconsistently** able to communicate well and build good rapport with patient and family, inconsistently demonstrates good relationship with colleagues, frequent fluctuation between “Below Expectations” and “Meets Expectations” performance, most importantly needs significant improvement before moving to a higher academic level	**Usually** able to communicate well with patient and family using appropriate terminology, usually good rapport in building a relationship with patient, family and/or team, usually sensitive to patient and colleagues	**Consistently** empathetic listening to patient, family and colleagues, excellent rapport with patients, families and team, patient centered communication, provides patient education, takes initiative in communicating with patient and colleagues

5. CRITICAL THINKING /CLINICAL DECISION MAKING. **NA □**

Please indicate the basis on which you are evaluating the student. Check all that apply.

Rounds □ Oral Presentations □ Written notes □ Precepting on clinical cases □ Other ……..

Below Expectations □	Needs Significant Improvement □	Meets Expectations Passably □ Sufficiently □	Exceeds Expectations Excellently □ Outstandingly □
Often unable to bring together history, physical exam and laboratory data , difficulty in interpreting data, problems lists are inaccurate or incomplete	**Inconsistently** able to use various clinical findings and data sources to develop a differential diagnosis, inconsistently comprehensive problem list, frequent fluctuation between “Below Expectations” and “Meets Expectations” performance, most	**Usually** able to synthesize various data sources, interprets basic data, develops differential diagnosis, problems list are usually complete	**Consistently** organizes various sources of data, correctly interprets data, develops coherent differential diagnosis and problem list, understands complex problems

	importantly needs significant improvement before moving to a higher academic level		

6. MANAGEMENT. **NA □**

Please indicate the basis on which you are evaluating the student. Check all that apply.

Rounds □ Conferences □ Presentations □ Precepting on clinical cases □ Other ……..

Below Expectations □	Needs Significant Improvement □	Meets Expectations Passably □ Sufficiently □	Exceeds Expectations Excellently □ Outstandingly □
Often unable to prioritize, analyze data and develop a plan, disorganized plan, often fails to consider patient's perspective, often lacks sound judgment	**Inconsistently** able to prioritize problems, develop a coherent management plan based on patient's problems and needs, frequent fluctuation between "Below Expectations" and "Meets Expectations" performance, most importantly needs significant improvement before moving to a higher academic level	**Usually** able to prioritize problems, develop a coherent plan for diagnosis and patient management, considers needs of patient, demonstrates sound judgment	**Consistently** prioritizes problems, demonstrates sound judgment, develops a coherent plan, includes patient and team in decision making, anticipates future problems

7. WRITTEN NOTES – Demonstrate ability to write history and physical examination, plans, progress note and discharge summary. **NA □**

Please indicate the basis on which you are evaluating the student.

Complete or focused history and physical □ Number of occurrences___ Progress notes □ Number of times___ Written Presentations □ Other ____

Below Expectations □	Needs Significant Improvement □	Meets Expectations Passably □ Sufficiently □	Exceeds Expectations Excellently □ Outstandingly □
Often incomplete, disorganized, tardy**,** unreliable, inaccurate, missing data, unrelated to problems, cursory notes	**Inconsistently** complete, organized and timely documentation, frequent fluctuation between "Below Expectations" and "Meets Expectations" performance, most importantly needs significant	**Usually** complete, organized, accurate on time, follow progress of patient, relate lab data to problems	**Consistently** comprehensive, organized, succinct, accurate, timely, relate lab data to problems

	improvement before moving to a higher academic level		

8. ORAL PRESENTATIONS – Demonstrate ability in patient presentations, conference presentations and formal student presentations. **NA □**

Please indicate the basis on which you are evaluating the student. Check all that apply.

Rounds □ Conferences □ Formal Student Presentations □ Precepting on clinical cases □ Other _____

Below Expectations	Needs Significant Improvement	Meets Expectations	Exceeds Expectations
□	□	Passably □ Sufficiently □	Excellently □ Outstandingly □
Often fails to present, inaccurate, poor organized**,** little evidence of preparation, not geared to audience, poorly understood concepts	**Inconsistently** organized, accurate and clear presentations, frequent fluctuation between “Below Expectations” and “Meets Expectations” performance, most importantly needs significant improvement before moving to a higher academic level	**Usually** organized, accurate, clearly presented, provides basic information, relates clinical problems to basic and clinical evidence, researched	**Consistently** organized, accurate, concise and clearly presented, well thought out, well researched, refers to evidence based medicine, demonstrates understanding of disease and management processes

9. SELF REFLECTION. **NA □**

Please indicate the basis on which you are evaluating the student.

Below Expectations	Needs Significant Improvement	Meets Expectations	Exceeds Expectations
□	□	Passably □ Sufficiently □	Excellently □ Outstandingly □
Often arrogant, antagonistic, fails to self-assess, blames others for failures, avoids taking responsibility, defensive, makes little effort to improve	**Inconsistently** demonstrates willingness to improve, inconsistently asks for feedback, frequent fluctuation between “Below Expectations” and “Meets Expectations” performance, most importantly needs significant improvement before moving to a higher academic level	**Usually** accepts full responsibility for actions, self-assesses, requests and improves with feedback, works toward self-improvement	**Consistently** accepts full responsibility for actions, self-assesses and recognizes strengths and weaknesses, seeks to improve performance, improves with feedback, self-directed, independent learner

10. PROFESSIONALISM. **NA □** Please indicate the basis on which you are evaluating the student. Check all that apply. Rounds □ Conferences □ Presentations □ Precepting on clinical cases □ Other ……..			
Below Expectations □	Needs Significant Improvement □	Meets Expectations Passably □ Sufficiently □	Exceeds Expectations Excellently □ Outstandingly □
Often unreliable, absent, late, fails to complete assignments, fails to accept responsibility, lacks motivation, records incomplete, disinterested, fails to maintain confidentiality, unethical behavior, handles stress poorly, lacks respect for patient, family, team, fails to recognize weaknesses	**Inconsistently** reliable, available, motivated and timely, inconsistently demonstrates respect toward patient and family, frequent fluctuation between “Below Expectations” and “Meets Expectations” performance, most importantly needs significant improvement before moving to a higher academic level	**Usually** reliable, present, on time, ethical, motivated, respects patient, family and team, works well with team, maintains confidentiality, follows through on tasks	**Consistently** reliable, active team member, respect of patients, families and team, ethical, maintains confidentiality, highly self-motivated, able to function independently but does to overstep bounds, follows through on tasks, flexible
11. SURGICAL SKILLS. **NA □** Please indicate the basis on which you are evaluating the student. Check all that apply. Direct Observation in OR, ER, clinic, bedside…□ Other (e.g. Skills lab) ……..□ This assessment encompasses the behavior of the student, including his/her ability to work efficiently, work with the surgical team, and particularly to facilitate *surgery*, that is, adapt and help, based on observation and instructions in the operating room. This assessment includes, but is not limited to, assessment of manual dexterity, the ability to demonstrate sterile technique, suturing and/or gluing simple lacerations, removal of sutures or staples, tying of surgical knots, demonstration of proper wound care, removal and application of dressing, including complex (packed) wounds, and other skills such as those listed in surgical encounters in Surgery Clerkship.			
Below Expectations □	Needs Significant Improvement □	Meets Expectations Passably □ Sufficiently □	Exceeds Expectations Excellently □ Outstandingly □

Often unable to complete basic surgical/wound care and similar technical exercises with basic skill or fails to improve skills due to lack of manual dexterity or unwillingness to respond to sustained guidance and support, fails to recognize weaknesses and/or does not work well in the operating room	**Inconsistently** able to complete surgical/wound care and similar technical exercises with basic skill or only modestly able to improve skills due to lack of manual dexterity; frequent fluctuation between “Below Expectations” and “Meets Expectations” levels, importantly needs improvement in skills and/or inconsistently works well in the operating room	**Usually** able to complete basic surgical/wound care and similar technical exercises with basic skill or improves skills demonstrating good manual dexterity, usually working well in the operating room with the team	**Consistently** able to complete basic surgical/wound care and similar technical exercises with basic skill or improves skills demonstrating exceptional manual dexterity, able to function independently but does to overstep bounds, follows through on tasks, flexible, works well in operating room with the team

Comments:

Evaluator’s signature:

Appendix C

Template Table Aligning Tests and Test Components with Outcome Competencies

Competency or Role	MCQ	OSCE							Clinical Performance Evaluation											Composite Score
		HT	PE	DS	CS	P	PS	MT	Kn	HT	PE	CS	DM	MT	WN	OP	SR	P	SS	
Scientist	✓✓								✓											CSc_Sc
Communicator		✓			✓✓					✓		✓✓			✓	✓				CSc_Co
Care Giver		✓	✓	✓			✓	✓	✓	✓	✓		✓	✓					✓	CSc_CG
Advocate								✓					✓	✓						CSc_Ad
Professional						✓✓											✓	✓✓		CSc_P

Abbreviations: MCQ: Multiple-Choice Questions; OSCE: Objective Structured Clinical Examination; HT: History Taking; PE: Physical Examination; DS: Diagnostic Skills; CS: Communication Skills; P: Professionalism; PS: Procedural Skills; MT: Management; Kn: Fund of Knowledge; DM: Decision-Making; WN: Written Notes; OP: Oral Presentations; SR: Self-Reflection; SS: Surgical Skills; CSc_Sc: Composite Score for Scientist Role; CSc_Co: Composite Score for Communicator Role; CSc_CG: Composite Score for Care Giver Role; CSc_Ad: Composite Score for Advocate Role; CSc_P: Composite Score for Professional Role.

Appendix D

Qualitative Analysis of Committee Deliberations

Theme	Category	Code	Frequency	Representative Quotations
Valuing	Professionalism	P	4	*"I looked especially at comments to identify students with problem in professionalism. I thought that professionalism cannot be rated well in the others ways."* *"I value professional behavior, and such looked at comments that describe professionalism."*
	Aggregated Skills	AS	2	*"I looked at each OSCE and checked every station separately to identify areas of weakness according to comments then to grades. I thought clinical skills are very important to look at first."* *"I dissected performance on OSCE because global skills matter a lot"*
Technical Process	Zoom-in	ZI	3	*"I started by looking at the final year grade. Then I thought that students who have scores between 70 & 75 are the students I need to know why they have low grades and then I went back to the rotations."* *"The end-of-year grade was a first step towards further consideration. Low performers were looked at separately and in details to make my selection."*
	Pattern recognition	PR	2	*"I looked at each final clerkship score for each student and I highlighted the weak students based on my previous experience. Then I read the comments about the highlighted students to select finally those with the bad/negative comments."*

				"Weak students can be identified from their borderline scores in everything"
	Probability building	PB	2	*"Since I am number-oriented person, I took the cut off value decided by the school of 70/100. So I did a formula to identify how many students have lowest than 70, then I highlighted those cases then I looked horizontally across all the grades. If I had the same pattern across clerkships, then I checked the comments."* *"I used the cut off of 70%. Those scoring above 70 on the final clerkship grades were considered eligible for promotion if they had succeeded in the OSCEs. Accordingly, I worked on a formula accounting for most scores but in a rank order."*
Validity	Accuracy	A	2	*"…there are too many discrepancies between the CPE and the comments. How can a student receive grades of 8-9 over 10 on an item and then get negative comments. So I did not depend on the CPE scores nor on the comments because I considered them inaccurate."* *"I don't consider CPE accurate; that is why I put more weight on the rest of the tests."*
	Discrimination potential	DP	2	*"…as I told you before the CPE did not reflect accurately students' performance and was not discriminatory (many items are rated all 8 and the narrative did not correlate with the grades) so I went to the OSCE, because I thought it was the most discriminatory."* *"MCQs and OSCEs are more discriminatory."*

VITA

NAME: Sola Aoun Bahous

EDUCATION: M.D., Lebanese University, Lebanon, 1995

M.S., Claude Bernard University, Lyon, France, 2000

Ph.D., Pierre and Marie Curie University Paris VI, France, 2005

TEACHING: Department of Basic Sciences, Balamand University, Faculty of Medicine and Medical Sciences, Koura, Lebanon: Pharmacology and Physiology for Medical Students, 2001-2008

Department of Basic Sciences, Lebanese American University School of Medicine, Byblos, Lebanon: Pharmacology and Physiology for Medical Students, 2008-present

Department of Medicine, Division of Nephrology, Lebanese American university Medical Center – Rizk Hospital, Beirut, Lebanon: Internal Medicine and Nephrology for Medical Students, 2010-present

HONORS: Cardiovascular Pharmacology, Pierre and Marie Curie University Paris VI, 2001-2005

PROFESSIONAL MEMBERSHIP:
Association of Medical Education in Europe (AMEE)
American Society of Nephrology
ERA-EDTA (European Renal Association)
Lebanese Society of Nephrology and Hypertension

ABSTRACTS: Aoun Bahous, S., Hariri, E., Mansour, A., Daaboul, Y., Korjian, S., El-Alam, A., Kilani, H., Karam, A., and Stephan, A.: Vitamin K2 supplementation and arterial stiffness in the renal transplant population – A single-arm, single center clinical trial. Poster presentation at the American Society of Nephrology Kidney Week Nov 2016.

Daaboul, Y., Korjian, S., El-Ghoul, B., Samad, S., Salameh, P., Dahdah, G., Hariri, E., Mansour, A., Spielman, K., Blacher, J., Safar, M.E., and Aoun Bahous, S.: Change in pulse wave velocity and short-term development of cardiovascular events in the hemodialysis population. *ATVB*, 36:A507, 2016.

Aoun Bahous, S., and Stephan, A.: Organ and tissue procurement as part of the curriculum of the medical and nursing schools. *Experimental and Clinical Transplantation,* L80, p. 96, 2014.

Bahous, S., and Stephan, A.: Aortic stiffness in living kidney donors. *Transplantation,*78(2):O152, p. 59, 2004.

PUBLICATIONS:

Safar, M.E., Gnakaméné, J-B., Aoun Bahous, S., Yannoutsos, A., and Thomas, T.: A longitudinal study of hypertensive subjects with Type 2 Diabetes Mellitus: Overall and cardiovascular risk. *Hypertension,* 69: 1029-1035, 2017.

El Ghoul, B., Daaboul, Y., Korjian, S., Khairallah, M., Karam, M.K., El Samad, S., Salameh, P., Dahdah, G., Blacher, J., Safar, M.E., and Aoun Bahous, S.: Impact of baseline kidney disease on arterial stiffness in end-stage renal disease: a cross-sectional study. *BioMed Research International*, Volume 2017, 2017, Article ID 2543262, https://doi.org/10.1155/2017/2543262.

Abi Raad, V., Raad, K., Daaboul, Y., Korjian, S., Asmar, N., Jammal, M., and Aoun Bahous, S.: Medical education in a foreign language and history-taking in the native language. *BMC Medical Education,*16: 298, 2016. DOI 10.1186/s12909-016-0826-7.

Korjian, S., Daaboul, Y., El-Ghoul, B., Samad, S., Salameh, P., Dahdah, G., Hariri, E., Mansour, A., Spielman, K., Blacher, J., Safar, M.E., and Aoun Bahous, S.: Change in pulse wave velocity and short term development of cardiovascular events in the hemodialysis population. *The Journal of Clinical Hypertension*, 18(9):857-863, 2016*.*

Aoun Bahous, S., Thomas, F., Pannier, B., Danchin, N., and Safar, M.E.: Country of birth affects blood pressure in the French hypertensive diabetic population. *Frontiers in Physiology,* 6: 1-7, 2015.

Korjian, S., Daaboul, Y., Stephan, A., and Aoun Bahous, S.: Organ procurement program: should we teach undergraduate medical and nursing students? *Experimental and Clinical Transplantation,*13: 55-58, 2015.

Aoun Bahous, S., Khairallah, M., Al Danaf, J., Halabi, R., Korjian, S., Daaboul, Y., Stephan, A., Blacher, J., and Safar, M.: Renal function decline in recipients and donors of kidney grafts: role of aortic stiffness. *American Journal of Nephrology,* 41: 57-65, 2015.

Yazbeck-Karam, V., Aoun Bahous, S., Faour, W., Khairallah, M., and Asmar, N.: Influence of standardized patient body habitus on undergraduate student performance in an objective structured clinical examination. *Medical Teacher,* 36(3): 240-244, 2014.

Mourad, J.J., Lopez-Sublet, M., Aoun Bahous, S., Villeneuve, F., Jaboureck, O., Dourmap-Collas, C., Denolle, T., Fourcade, J., and Baguet, J.P.: Impact of miscuffing during home blood pressure measurement on the prevalence of masked hypertension. *American Journal of Hypertension,* 26(10): 1205-1209, 2013.

Aoun Bahous, S., Stephan, A., Blacher, J., and Safar, M.: Cardiovascular and renal outcome in recipients of kidney grafts from living donors: role of aortic stiffness. *Nephrology Dialysis Transplantation*, 27: 2095-2100, 2012.

Aoun Bahous, S., Blacher, J., and Safar, M.E.: Aortic stiffness, kidney disease, and renal transplantation. *Current Hypertension Reports,*11: 98-103, 2009.

Safar, M.E., Delahousse, M., and Aoun Bahous, S.: Arterial stiffness and renal transplantation. *Journal of Hypertension*, 26: 2101-2102, 2008.

Aoun Bahous, S., Stephan, A., Blacher, J., and Safar, M.E.: Aortic stiffness, living donors and renal transplantation. *Hypertension*, 47: 216-221, 2006.

Aoun Bahous, S., Stephan, A., Barakat, W., Blacher, J., Asmar, R., and Safar, M.E.: Aortic pulse wave velocity in

renal transplant patients. *Kidney International*, 66: 1486-1492, 2004.

Barbari, A., Stephan, A., Masri, M., Karam, A., Aoun, S., El Nahas, J., and Bou Khalil, J.: Consanguinity-associated kidney diseases in Lebanon: an epidemiological study. *Molecular Immunology*, 39: 1109-1114, 2003.

Ramos, E., Aoun, S., and Harmon, W.E.: Expanding the donor pool: effect on graft outcome. *J Am Soc Nephrol.* 13(10): 2590-2599, 2002.

Aoun, S., Blacher, J., Safar, M., and Mourad, J.J.: Diabetes mellitus and renal failure: effects on large artery stiffness. *Journal of Human Hypertension*, 15(10): 693-700, 2001.

Aoun, S., and Ramos, E.: Hypertension in the HIV-infected patient. *Current Hypertension Reports,* 2(5): 478-481, 2000.

Aoun, S., and Ramos, E.: Expanding the donor pool: effect on graft outcome. *Transplant Proc.* 31(8): 3379-82, 1999.

Masri, M.A., Barbari, A., Stephan, A., Kamel, G., Aoun, S., Rizk, S., and Karam, A.: Safe and cost effective conversion from Neoral to Consupren soft gelatin capsules in stable renal transplant patients: a 1-year study. *Transplant Proc.* 31(8): 3302-3, 1999.

Stephan, A., Barbari, A., Kamel, G., Aoun, S., and Masri, M.A.: A one-center experience with a short course of mycophenolic acid. *Transplant Proc.* 31(8): 3289-3290, 1999.

Stephan, A., Masri, M.A., Barbari, A., Aoun, S., Rizk, S., and Kamel, G.: A one-year comparative study of Neoral vs Consupren in de novo renal transplant patients. *Transplant Proc.* 30(7): 3533-4, 1998.