Original Article

Access this article online
Quick Response Code:
Website: www.jehp.net
DOI: 10.4103/jehp.jehp_1500_20

Providing a model for validation of the assessment system of internal medicine residents based on Kane's framework

Mostafa Dehghani Poudeh, Aeen Mohammadi[1], Rita Mojtahedzadeh[1], Nikoo Yamani[2], Ali Delavar[3]

Department of Medical Education, School of Medicine, Tehran University of Medical Sciences, Tehran, Iran, [1]Department of E-learning in Medical Education, Virtual School, Center for Excellence in E-learning in Medical Education, Tehran University of Medical Sciences, Tehran, Iran, [2]Department of Medical Education, Isfahan University of Medical Sciences, Isfahan, Iran, [3]Department of Evaluation and Measurement, School of Education and Educational Psychology, Allameh Tabatabaei University, Tehran, Iran

Address for correspondence:
Dr. Aeen Mohammadi, Department of E-learning in Medical Education, Virtual School, Center for Excellence in E-learning in Medical Education, Tehran University of Medical Sciences, Tehran, Iran.
E-mail: aeen_mohammadi@tums.ac.ir

Received: 28-11-2020
Accepted: 29-03-2021
Published: 29-10-2021

Abstract:

BACKGROUND: Kane's validity framework examines the validity of the interpretation of a test at the four levels of scoring, generalization, extrapolation, and implications. No model has been yet proposed to use this framework particularly for a system of assessment. This study provided a model for the validation of the internal medicine residents' assessment system, based on the Kane's framework.

MATERIALS AND METHODS: Through a five stages study, first, by reviewing the literature, the methods used, and the study challenges, in using Kane's framework, were extracted. Then, possible assumptions about the design and implementation of residents' tests and the proposed methods for their validation at each of their four inferences of Kane's validity were made in the form of two tables. Subsequently, in a focus group session, the assumptions and proposed validation methods were reviewed. In the fourth stage, the opinions of seven internal medicine professors were asked about the results of the focus group. Finally, the assumptions and the final validation model were prepared.

RESULTS: The proposed tables were modified in the focus group. The validation table was developed consisting of tests, used at each Miller's pyramid level. The results were approved by five professors of the internal medicine. The final table has five rows, respectively, as the levels of Knows and Knows How, Shows How, Shows, Does, and the fifth one for the final scores of residents. The columns of the table demonstrate the necessary measures for validation at the four levels of inferences of Kane's framework.

CONCLUSION: The proposed model ensures the validity of the internal medicine specialty residency assessment system based on Kane's framework, especially at the implication level.

Keywords:

Educational measurement, graduate, internship and residency, Kane's framework, medical, reliability and validity, validity of results

Introduction

Having a continuous, comprehensive and developed assessment and feedback system, mainly in the clinical environment, is one of the requirements of competency-based medical education.[1] In fact, in order to evaluate learners, this system must use various methods and tools, both quantitative and qualitative, as well as formal and informal (such as observations and subjective assessments during the course) in accordance with the evaluated competencies in both formative and summative approaches, and provide feedback to learners.[2,3] Therefore, since in the new systems of assessments, the emphasis on scores and the use of benchmark-reference tests at the same time, has given way to the use of multiple

How to cite this article: Poudeh MD, Mohammadi A, Mojtahedzadeh R, Yamani N, Delavar A. Providing a model for validation of the assessment system of internal medicine residents based on Kane's framework. J Edu Health Promot 2021;10:386.

and varied methods of assessment throughout the course, using only traditional methods of determining validity (content, criteria, or structure) will provide limited results. Therefore, this system should use appropriate and up-to-date theories and frameworks for validation.[4] Furthermore, according to Harris *et al*., validity is not a number but an argument. Therefore, the frameworks used to ensure validity must be able to answer complex questions about interpretations of scores and the alignment of these interpretations with theories and observations. In addition, as the use of a single tool is not sufficient to measure even one competency and different methods and tools must be used, even the validity of each of these individual tools would lead to no desired result. Rather, it is how they are combined in the form of an assessment system that must be examined.[5] In other words, in addition to evaluating the quality of the tools, the overall validity of an assessment program or system must be assessed using sufficient evidence. On the other hand, the use of new methods and approaches to quality control of learners' assessment is considered as a requirement for modern validation systems.[6] In this regard, what has been used in recent studies in the field of test validity is Kane theory of validity.[7-9] He suggested that four inferences should be considered to ensure the validity of an assessment.

1. In the first inference, it should be determined whether the scores, obtained from observing the learners' performance, had necessary accuracy or not, and what evidence and documents can be collected and presented to prove this claim? This inference is called scoring
2. This author believes that then, there should be evidence of the generalizability of the results to the total expected results of universe score. He called this inference as generalization
3. The third inference is called extrapolation. At this inference, the possibility of using scores to infer about the expected competencies of the assessed in the practical environment is examined
4. Finally, at the last inference or the implications, the correctness of the judgments, made about the ability of learners and the decision to allow them to enter the professional field of work and activity, is examined.

Experimental studies on the validity of assessment systems indicate that despite the appropriate diversity of assessment methods, few of these studies have used Kane's framework.[7] In some studies, only a part of the assessments has been validated using this framework.[10-15] However, Wools, Eggen and Béguin have used this model to determine the validity of assessments during social workers' training.[16] In addition, in various studies, not all four inferences of Kane's Framework have been considered equally, and in some articles that have provided recommendations for the use of this framework, the recommendations have not been the same for all inferences. For example, Cook *et al*. had no specific proposal for the implications inference.[17] However, although frameworks for validation studies have been proposed in recent years, and despite conceptual developments,[5,17] this theory still needs further elaboration for the implementation as well as simplification for users. Some authors have even suggested that this requires the use of a comprehensive measure to determine the validity of all the tools, used in making decisions.[16,18]

However, ensuring the validity of the assessment of specialty programs of medical learners whose graduates are allowed to enter professional fields to practice medicine independently will be of particular importance. Therefore, the assessment of the abilities of this group of learners should be done through a coherent system with various methods. One of the mother disciplines, which includes several sub-disciplines, is the field of internal medicine. Therefore, it is very important to define and determine methods and measures that can guarantee the quality of the assessment system of specialty residents in this field. Therefore, as in most studies, only instruments and tests have been studied for validity individually, and the validity of the assessment system has never been examined. In this paper, a practical model for implementing Kane's framework to evaluate the validity of the assessment system of internal medicine residents is proposed. This model can also be used to ensure the validity of the assessment system of other specialty programs.

Materials and Methods

Study design and setting

This multimethod study was performed in 2020 in five stages in the Department of Internal Medicine, Isfahan University of Medical Sciences. Internal medicine residents are evaluated through an assessment system consisting of a variety of assessment methods including multiple-choice written and descriptive tests, objective structured clinical examination (OSCE), mini clinical evaluation examination (mini-CEX), PMP, logbook, professors' overall scores on monthly rotations in the internal medicine subspecialties, professional behavior score, 360° test, and performance quality score in hospital wards. The final score of each resident at the end of each year is the sum of the scores of each of the above assessments. Not all methods have the same weight and value, and each has a specific percentage of the final score. In addition, this percentage varies between different years of residency. For example, scoring in the second year of residency is composed of the minimum score of ward (40) + OSCE (30) + mini-CEX (30) + professional behavior (30) + file recording (10) and + logbook score.[10]

The minimum score of the ward in the 1st year is 30 points. Furthermore, the minimum total score must be 150. Finally, the final score is only 50% of each residency year-end score. The rest of the final annual score is obtained from the promotion test (progress test), which is held annually in a centralized format.

Study participants and sampling

The study was conducted on the assessment system of the residents of the general specialty of internal medicine.

Data collection tool and technique

A systematic review, developing the proposed assumptions of the tests and their validation methods, conducting a focus group with internal medicine professors, confirming the results by additional professors of the internal medicine department and developing the final model were the corresponding stages.

Ethical consideration

Prior to the meeting, informed consent was obtained verbally from the participants to participate in this issue. The study was approved by the Tehran University of Medical Sciences research ethics committee.

First, through a systematic review of existing literature in the field of medical education, the methods used until the end of year 2020 in validation, based on Kane's framework, as well as the challenges reported in these studies were extracted. This review was conducted on the Web of Science, Scopus, Pub Med, Embase, Science Direct, and Ovid databases. The keywords used were as follows:

Student' learner' medical student undergraduate' valid*' content validity' validity theory' validity assessment, Kane's validity, Kane's framework, and Kane's theory.

As per the inclusion and exclusion criteria, in this review, only articles that were used to determine the validity of medical students' assessment using Kane's theory were reviewed. There was no time limit and only articles, published in English, were included in the study. A variety of review articles and empirical studies were reviewed. Therefore, studies in nonmedical fields and in non-English language, either in which other validity models rather than Kane's framework were used or in which other Kane's theories were used, were excluded from the study. Since our goal was to use Kane's model in evaluating the validity of academic achievement tests, studies in the field of postgraduate education (continuing education in the medical community) were also excluded from this review. Figure 1 shows the steps of this search.

Then, in order to determine the method of examining the desired interpretations in each of the internal medicine tests, the assumptions, considered by the designers and organizers of these tests, were extracted in the main and sub-category inferences, and at four Kane's levels. These assumptions were first compiled by the first author based on available sources on Kane's framework, and then, finalized in a meeting with members of the research team. At each inference, each of the main assumptions contained a sentence or paragraph that specified the overall purpose and end result of the validation of that inference level. However, the sub-category assumptions were the assumptions that would be the criterion for proving or rejecting the main assumptions, and finally, the interpretation of the test in question.

In the next stage, a focus group meeting, consisting of members of the research team, the head of the department, professors in charge of training and assessment of residents of the internal medicine department and three other professors of this department, selected by the head of the department, was held. In addition to recording the audio, the content was recorded by the secretary during the session. The meeting lasted 3 h. At the beginning of the session, explanations were given about Kane's validity and framework. Participants were then asked to read the proposed main and sub-assumptions of each inference individually for 5 min first. After the reading, the participants expressed their views on the assumptions. The Kane inferences (first the main assumptions and then the sub-assumptions of each inference) were then discussed and rewritten. At the end of this section, the final main and sub-assumptions were identified through voting and if necessary, through the consensus. These assumptions formed the basis for determining the statistical methods, arguments, and documentation, needed to determine the validity of each of the department tests at Kane's four levels of inferences. In the second part of the focus group meeting, the methods and documents, used in other studies, which had been prepared in the form of a table in advance, were provided to the members of the meeting. The members of the focus group were then divided into two groups. In each group, the proposed methods and documentation for the two inferences of Kane's framework (based on the final assumptions) were reviewed. Then, the results of the studies were presented in a joint session and as in the first part, were finalized. The result was a practical model for assessing the validity of the assessment system of internal medicine residents.

In the next step, the final model was sent through E-mail to seven other professors in the department who participated in training and assessment of residents but was not members of the focus group to express their suggestions for finalizing the results. These academic members were introduced to the research team by the person in charge of evaluating the internal medicine

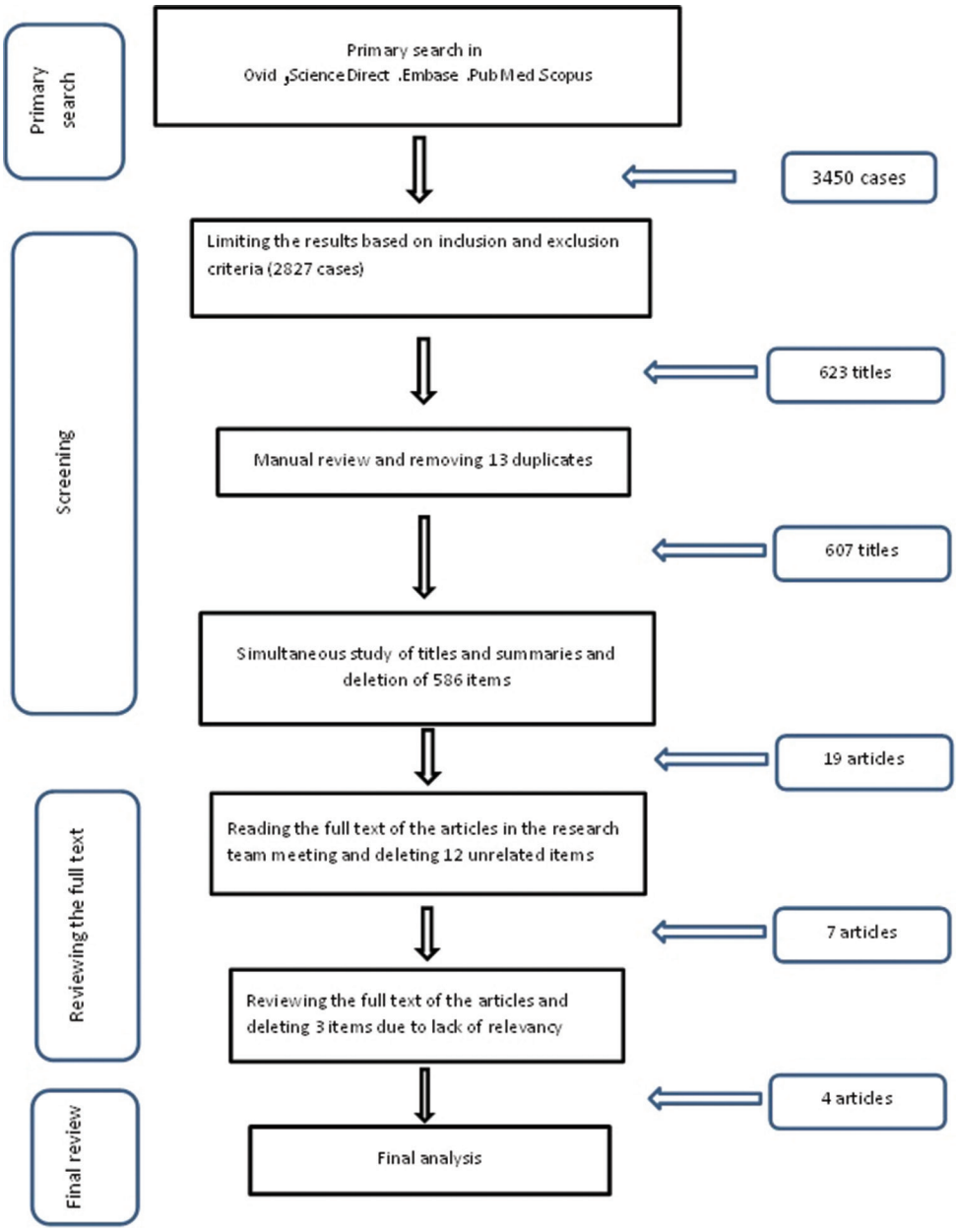

Figure 1: Flowchart of systematic review of Kane's validity studies

residents. In the E-mail sent, the professors were asked to check the importance and feasibility of each of the main and sub-assumptions and validation methods, and if not possible, to state in writing the reason, as well as their suggested alternative method.

In the last step, the response to E-mails and the results of the previous steps were examined in the research team meeting. In cases where the response to the E-mail needed further explanation, the necessary explanation was obtained from the relevant professor by telephone at the same meeting by the first author of the article. Thus, the final assumptions and validation model were determined based on Kane's framework. Figure 2 summarizes the study process.

Results

In the first phase of this study, as shown in Figure 1, in the first round after the initial search in the databases, 3450 titles were obtained, and after removing 2827 titles according to inclusion and exclusion criteria, 623 items were found, of which 13 were related to web of science, 180 to Scopus, 27 to PubMed, 400 to science direct publications and 11 to Ovid. A search on the Embase database returned no results. These findings were entered into Endnote X9 software and reviewed and screened. After reading the full text of the selected articles, three more articles were deleted as one of the articles contained only the opinions and suggestions

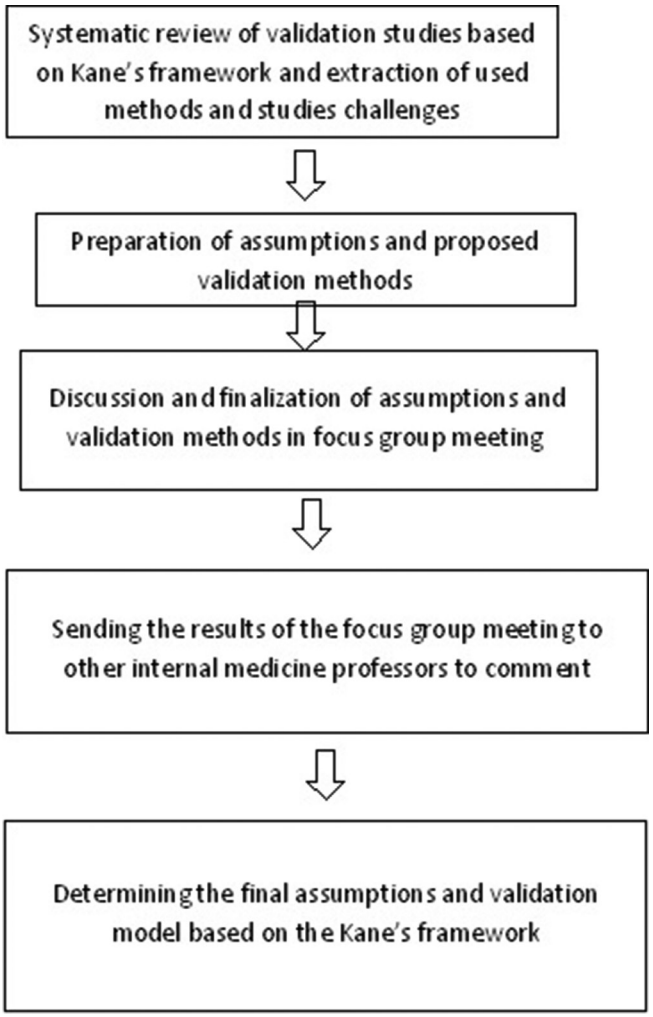

Figure 2: Flow chart of the study method

of its author,[15] the other article examined and proved three claims about portfolio in dentistry by referring to Kane's framework.[19]

The third paper used Kane's composite reliability framework.[20] A review of studies conducted on Kane's framework showed that in general, at the scoring inference, items related to the preparation and use of tools and methods of assessment and their quality, training and debriefing of evaluators, scoring processes, and scoring distribution patterns can be enumerated. At the inference of generalization, the quality of different sampling methods, the agreement between the evaluators, and the results of different assessments and different reliability calculations will be included. In validity at the inference of extrapolation, which had the highest frequency among different methods, using different methods such as checking the ability to differentiate between different levels of participants, coherence, and consistency between the assessments performed in different educational stages or different tests of an individual stage can be reviewed. Finally, in this study, we faced a significant lack of evidence at the implication/decision level.

The assumptions extracted in the second step of the research can be seen in the form of Appendix 1 of this article. At this stage, four main assumptions and 21 sub-assumptions were prepared. Of these, seven sub-assumptions were related to the scoring inference, 6 sub-assumptions were related to the generalizability inference, 4 sub-assumptions were related to the extrapolation inference, and 4 sub-assumptions were related to the decision level. Suggested methods for validating the Residents' Assessment System are also given in Appendix 2.

In the third step (in the focus group meeting), four main assumptions and 21 sub-assumptions were obtained. Of these, 8 sub-assumptions were related to the scoring, 5 were related to the generalizability, 4 were related to the extrapolation, and 4 sub-assumptions were related to the decision inference level.

In the fourth step, five professors responded to the E-mail and announced their comments in the submitted tables.

In the last step, based on the results of the previous steps, the assumptions and the final table of validation methods of the residents' assessment system were prepared. Thus, the final assumptions were obtained according to Table 1. However, in brief, these assumptions can be described as follows:

- Assumptions of the scoring inference were about the design and conduct of regular and correct tests as well as the quality assurance methods of the questions and observations
- Assumptions of the inference of generalization took into account the assurance of proper sampling and coverage of the course content and the generalizability of the assessment results
- Extrapolation-inference assumptions were about ensuring that test results were true and that the tests were logically related
- At the implications inference, the final assumptions were about the department's confidence in the correctness of the decisions, made about the residents.

Finally, the final model, obtained in the fifth and final step shown in Table 2, is the final methods and measures that are used at different levels of competence as well as the four Kane's inferences in the validation of the residency assessment system. As can be seen in this table, the tests were also categorized based on the Miller pyramid and placed in the first column of the table. The other columns were the test type and Kane's quadruple inference levels. The tests of this system were also given in the rows of this table.

Table 1: Assumptions in the assessment system of internal medicine assistants at the four inferences of Kane's validity framework

Validity level	Assumptions	Sub-assumptions
Scoring	The test is properly designed and executed, and also, the scores are a true and accurate representation of the observations. In other words, in addition to the fact that the observations must be done according to the principled and correct methods, the translation of the observations into scores has also been done correctly	The designers of the various test questions have received the necessary training on the characteristics of each test
		The assessment system has a comprehensive plan and an overall blueprint
		Tests cover different inferences of competence
		The schedule of tests has gone according to the plan
		Each test has a blueprint and the questions are formulated accordingly
		The minimum passing score in the assessment system under the study is determined based on coherent and logical methods and based on the scientific principles
		Each of the tests has good internal consistency
		The design of the questions, the holding and execution of each test has proceeded according to scientific principles
Generalization	The tests evaluate appropriate examples of the competencies, expected from the residents, and their results can be generalized to all competencies	Tests are a good example of the different levels of Miller Pyramid competencies
		Test items are a good example of the content to be evaluated
		The tests have good reliability (error rate is low)
		The difference between the scores in the tests in general is just due to the real difference between the abilities of the residents and not due to other factors
		Residents' final tests and scores have an acceptable generalizability coefficient
Extrapolation	In addition to being correlated with each other, tests of different inferences of competence also have good predictability for each other	The test results are such that they distinguish the residents of the older years from the residents of the younger years.
		The questions, scenarios, and problems of the patients, raised in the tests, correspond to the real-world conditions
		There is a good correlation between the scores of the corresponding competencies in different tests
		Low levels of competency scores predict higher levels
Implications	Granting assistants to enter promotion and board exams is consistent with their actual performance throughout the year in the workplace and the results of promotion and board exams	There is a correlation between the scores obtained in the group exams, and the score of the regional or national promotion exam and board exam
		The scores, obtained in the group exams, are correlated with the general opinions of the professors about each assistant
		Test scores show the trend of increasing the experience and ability of residents in each year
		Test scores show the trend of increasing the experience and ability of residents during the course in 1 year

Discussion

The aim of this study was to obtain a practical model for implementing Kane's framework in order to determine the validity or validation of an assessment system in the field of internal medicine. In fact, what was obtained as a result of the present study, through the results of other studies as well as the agreement of the professors involved in the assessment of internal medicine residents, will be a good guide for researchers and evaluators to evaluate and improve the quality of learners' assessment systems in form of a checklist that can be the basis for ensuring the validity of the assessment system. The advantage of this model over similar cases is that it determines in detail the validation of each of the tools and the methods of assessment of residents, based on Kane's framework of methods, actions and the documentation required, while in other suggested cases, it was not so. For example, Cook *et al*. focused only on the inferences of this framework and recommended items for each inference and did not pay attention to assessment methods and tools.[17] In addition, in this model, special attention is paid to the competencies of residents and their use as a basis for validation by gathering the necessary documents and evidence. This is important because according to the new approach to assessment, each competency is evaluated using different methods and tools. Therefore, according to the recommendation of the initiators of this approach, measuring the validity and reliability of assessments should prove the convergence of these methods in evaluating competencies,[3,5] and this is the point that is well addressed in the proposed model of the present study.

Moreover, its application in both medical and nonmedical sciences indicates that this framework is a model that can well assess the validity of different assessment tools or methods. But as Kane himself acknowledges, determining validity based on interpretations and intended uses of tests and scores is not an easy task due to the load and burden of work.[20] Therefore, according to this author's recommendation, instead of examining the system components, it is better to examine the items that are more questionable and more important

Table 2: Tools and methods used to evaluate the validity of the system of assessment of specialized assistants for internal diseases based on the Kane's framework

Competency level	Type of test	Methods and measures required to validate tests at each level of the Kane's framework			
		Scoring	Generalization	Extrapolation	Implications
Knows and knows how	Written exams, multiple choice, descriptive tests and PMP[a]		Reviewing the results of test analysis Checking the status of sampling questions in blueprint Checking the generalizability coefficient[b] of tests	Investigating the difference in scores in different years of residency Checking the authenticity of the scenarios[c] Checking the correlation of the corresponding questions in different tests	Checking the correlation of scores with the results of the progress test[d] Assessing the correlation between test/assessment results based on the general opinion of professors[e]
Shows how	OSCE	Checking the quantity and quality of training of question designers Checking how to prepare station checklists Checking the quantity and quality of observer's training Completing of the test quality checklist Checking how the stations are arranged for each year of residency Checking how to determine the standard setting of each year at each station	Examining how the curriculum is sampled to determine stations Testing the reliability of the test[f] Investigating the correlation between the scores of the residents of the parallel lines of the test and different times Checking the sources of error in the test Checking the test generalizability coefficient	Investigating the correlation between station scores and corresponding tests[g] Checking the authenticity of the scenarios Checking the difference in grades in different years of residency	Checking the correlation of scores with progress test results Assessing the correlation between test/assessment results based on the general opinion of professors
Shows	Mini-CEX	Checking the quantity and quality of training of question designers Completing the quality checklist of the exam Ensuring that the assessor's gender is not related to the score Checking the strictness and lenience of professors Assessing the satisfaction of residents and evaluators of the test	Frequency of test components (patient type, test setting, disease complexity, test focus type) Investigating the effective factors in the variation of test scores Checking the test generalizability coefficient	Reviewing the progress of scores in different months Investigating the correlation of corresponding competencies in different tests	Investigating the correlation of scores with the results of the progress test Assessing the correlation between test/assessment results based on the general opinion of professors Reviewing the results of feedbacks
Does	Intra-wards score and 360-degree assessment	Checking the holding according to the comprehensive schedule of residents' exams Checking how to complete the tool	Checking the reliability of scores	Checking the correlation between the corresponding items in different tests	Assessing the correlation between test/assessment results based on the general opinion of professors
	Professional behavior score	How to compile test tools Checking how to complete the tool	Checking the reliability of scores	Checking the correlation between the corresponding items in different tests	Checking the correlation between the corresponding items in different tests
	Logbook	Checking how the residents complete the logs Checking how the teachers score	Checking the reliability of scores	Checking the correlation between the corresponding items in different tests	Assessing the correlation between test/assessment results based on the general opinion of professors
	Record writing score	How to design test tools Checking how to complete the tool	Checking the reliability of scores	Checking the correlation between the corresponding items in different tests	Assessing the correlation between test/assessment results based on the general opinion of professors

Contd...

Table 2: Contd...

Competency level	Type of test	Methods and measures required to validate tests at each level of the Kane's framework			
		Scoring	Generalization	Extrapolation	Implications
Final scores		Checking the conformity of how to calculate the score with the regulations	Investigating the factors of variations in scores between assistants Checking the reliability of scores Checking the total generalizability coefficient	Checking the correlation between the corresponding items in different tests	Investigating the effect of residency year on scores Evaluating the correlation between the final score and the assessment based on the general opinion of the professors

[a]Patient Management Problems is a written test to assess problem-solving ability or clinical reasoning, [b]The purpose is to statistically calculate the degree of generalizability of the results to the total expected results of the examinee, [c]The first part of each question in the medical exams, which describes the main situation and context of the problem to ask the relevant questions, [d]It is a written test that is held at the end of each residency year to grant entry permission to a higher year, [e]Comments that are made at the end of each year by the professors on a subjective basis about each resident, [f]The reliability of a test shows the degree of reproducibility of scores or test results and is calculated by determining the degree of correlation between the scores obtained from the repetition of a test or two halves of a test, [g]Corresponding tests or competencies are tests or competencies that measure a common construct. OSCE=Objective Structured clinical Examination, Mini-CEX=Mini Clinical Evaluation Examination, PMP=Patient Management Problems is a written test to assess problem-solving ability or clinical reasoning.

than other parts of the assessment system.[21] This not only avoids wasting time, energy and resources, but also prevents complexity of the analysis and facilitates the final conclusions. As in the study of validity, Bok *et al*. focused only on the inference of generalization of test the results.[22] In this regard, the application of the proposed methods in the present study, especially common methods, prevents redundancy in the analysis of individual tests as some of these methods can be used on more than one or two inferences. Kelly-Riley and Elliott have proposed a review of the reliability and coherence of scores for the scoring inference[23] while it seems that these methods also evaluate the inference of generalization. Moreover, there are cases that have been used innovatively in some individual studies or in a specific assessment method.[24]

But there are two notable points in the results of the present study. The first is to consider the competency levels, proposed by Miller in the test category[25] and to determine the assumptions and possible uses of them at each inference of Kane's framework, and consequently, how to assess the extent to which these assumptions are met. In other words, the product of the actions taken and the interpretation of the documents collected at each of the four inferences of Kane must be to confirm or reject the assumptions of the intended inference. For example, examining the quantity and quality of question designers' training can show whether the question designers have undergone various tests designing courses on the characteristics of each test. This point, along with the documentation of other sub-assumptions at this inference, such as designing questions, holding and conducting any test according to scientific principles, will generally confirm or contradict the design and conduct of regular and correct tests, as well as designing quality questions and observations have been. But, it is important to consider Miller's inferences of competence because the validity of an assessment system depends on choosing the right tools at each level of the pyramid.[26] This can be achieved by having a comprehensive assessment program as well as a test blueprint, both of which are considered in the proposed model of this study. The second point is the specific proposition for validity argument at the fourth inference, the implications inference, as many validation studies with Kane's framework have not been very successful in providing the necessary documentation for this inference.[27] What is proposed in the present study is to compare the general opinions of professors about the resident in performing the assigned tasks with the results of different tests; comments that are the result of the collective agreement of the evaluators. To this end, Entrustble Professional Activities (EAPs) can be a good tool and provide reliable results. These activities are by definition the core tasks of a discipline (profession, specialty, or subspecialty) that a person can perform reliably without direct supervision in a specific environment providing health services after demonstrating sufficient competence.[28,29]

These activities are used by the professors during the residency period as a basis for measuring the level of capability of the residents in the relevant specialized fields. Although these activities have been proposed for formative use, their use in a specific period of time and even as a basis for final decisions has been suggested.[30] Therefore, the assessment of the capability of residents to perform specialized tasks, assigned at the end of each year of the residency, can be used as evidence to check the validity of the results of assessments, conducted during the year. It should be noted that the EPA is different from Milestone. In fact, the EPA is a task in the field of medical science that must be performed by a learner at a level of need for supervision. Milestone, on the other hand, is an individual trait or attribute that the learner must acquire at some point in time.[30-33]

For example, taking a history and performing a physical exam is a task or action that is covered by the EPA.

The corresponding milestone, on the other hand, is the fit of the history taken, with the patient's condition or its accuracy, as well as the taking such a history that provides the information, needed for the diagnosis.[29] As can be seen, in the latter, there is talk of a feature necessary for taking a history. In this way, it can be judged whether the decisions made by the training group using the results of practical tests (i.e., mini-CEX or OSCE) are consistent with the assistants' skill status in performing their duties. Of course, the allocation of this situation should be determined by creating a consensus about the mentioned task among the professors who have had enough encounters with the resident to be evaluated.

Conclusion

Finally, it should be noted that although the present study was conducted on the assessment system of internal medicine residents, its results can be well applied in many systems and even individual assessment tools of medical programs, particularly in non-surgical residency assessment systems. Meanwhile, further research, in addition to expanding the results of the present study, can be used to determine and ensure the validity of a comprehensive assessment system (programmatic assessment). However, the assumption of promoting residents' learning has been seen with the methods, mentioned in this model, such as determining the effect of the year of the residency on scores or examining the progress of scores in different months.

Limitation and recommendation

One of the limitations of this study was that it was conducted in one university and did not include the opinions of other universities, which could be the next step in this study. Furthermore, the evidence and documents included in our proposed model, on the one hand, have been used jointly in various studies, and on the other hand, their collection has been confirmed by the professors of the internal medicine department. Therefore, it is suggested that the next step in the continuation of this research be to formulate and use the EPAs of internal medicine residents to check the validity of the implications inference. In addition, our proposed model should be used in practice to be validated and its shortcomings to be identified to be used in future studies of the validity of assessment systems. Finally, by reflecting on the findings of the final model, the actions and documents that can be used jointly to determine the validity of different inferences and tests during the implementation of the model can include examining issues such as blueprints, designing quality and delivering various tests, correlation between exams, correlation between the questions of each exam, correlation between the corresponding abilities in different exams, correlation of scores with general opinions of professors about each assistant, and finally, the congruence of test results to the years of residency and also to the passage of time in each year.

Acknowledgment

This article is part of the results of the first author's doctoral thesis with ethics code IR.TUMS.MEDICINE.REC.1399.1047, registered in Tehran University of Medical Sciences.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Van Der Vleuten CP, Schuwirth LW, Driessen EW, Govaerts MJ, Heeneman S. Twelve tips for programmatic assessment. Med Teach 2015;37:641-6.
2. Lockyer J, Carraccio C, Chan MK, Hart D, Smee S, Touchie C, *et al*. Core principles of assessment in competency-based medical education. Med Teach 2017;39:609-16.
3. Schuwirth LW, Van der Vleuten CP. Programmatic assessment: From assessment of learning to assessment for learning. Med Teach 2011;33:478-85.
4. Harris P, Bhanji F, Topps M, Ross S, Lieberman S, Frank JR, *et al*. Evolving concepts of assessment in a competency-based world. Med Teach 2017;39:603-8.
5. Schuwirth LW, van der Vleuten CP. Programmatic assessment and Kane's validity perspective. Med Educ 2012;46:38-48.
6. van der Vleuten CP, Dannefer EF. Towards a systems approach to assessment. Med Teach 2012;34:185-6.
7. Im GH, Shin D, Cheng LJ. Critical review of validation models and practices in language testing: Their limitations and future directions for validation research. Language Testing in Asia 2019;9:14.
8. Kane MT. Validating the interpretations and uses of test scores. Journal of Educational Measurement. 2013;50:1-73.
9. Kane MT. Validation as a pragmatic, scientific activity. Journal of Educational Measurement. 2013;50115-22.
10. Tavares W, Brydges R, Myre P, Prpic J, Turner L, Yelle R, *et al*. Applying Kane's validity framework to a simulation based assessment of clinical competence. Adv Health Sci Educ Theory Pract 2018;23:323-38.
11. Bajwa NM, Yudkowsky R, Belli D, Vu NV, Park YS. Improving the residency admissions process by integrating a professionalism assessment: A validity and feasibility study. Adv Health Sci Educ Theory Pract 2017;22:69-89.
12. Hatala R, Cook DA, Brydges R, Hawkins R. Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS): A systematic review of validity evidence. Advances in health sciences education: Theory and practice. 2015;20:1149-75.
13. Clauser BE, Margolis MJ, Holtman MC, Katsufrakis PJ, Hawkins RE. Validity considerations in the assessment of professionalism. Adv Health Sci Educ Theory Pract 2012;17:165-81.
14. Surry LT, Torre D, Durning SJ. Exploring examinee behaviours as validity evidence for multiple‐choice question examinations. Medical education 2017;51:1075-85.
15. Peeters MJ, Martin BA. Validation of learning assessments:

A primer. Curr Pharm Teach Learn 2017;9:925-33.
16. Wools S, Eggen TJ, Béguin AA. Constructing validity arguments for test combinations. Studies in educational evaluation. 2016;48:10-8.
17. Cook DA, Brydges R, Ginsburg S, Hatala RJ. A contemporary approach to validity arguments: A practical guide to Kane's framework. Medical education. 2015;49:560-75.
18. Johnson RC, Riazi AM. Validation of a locally created and rated writing test used for placement in a higher education EFL program. Assessing Writing. 2017;32:85-104.
19. Gadbury-Amyot CC, McCracken MS, Woldt JL, Brennan RL. Validity and reliability of portfolio assessment of student competence in two dental school populations: A four-year study. J Dent Educ 2014;78:657-67.
20. Onishi H, Park YS, Takayanagi R, Fujinuma Y. Combining scores based on compensatory and noncompensatory scoring rules to assess resident readiness for unsupervised practice: Implications from a national primary care certification examination in Japan. Acad Med 2018;93:S45-51.
21. Kane MT. Explicating validity. Assessment in Education: Principles, Policy, & Practice. 2016;23:198-211.
22. Bok HG, de Jong LH, O'Neill T, Maxey C, Hecker KG. Validity evidence for programmatic assessment in competency-based education. Perspect Med Educ 2018;7:362-72.
23. Kelly-Riley D, Elliot NJ. The WPA outcomes statement, validation, and the pursuit of localism. Assessing writing. 2014;21:89-103.
24. Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. Adv Health Sci Educ Theory Pract 2014;19:233-50.
25. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990;65:S63-7.
26. Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet 2001;357:945-9.
27. Hawkins RE, Margolis MJ, Durning SJ, Norcini JJ. Constructing a validity argument for the mini-Clinical Evaluation Exercise: A review of the research. Acad Med 2010;85:1453-61.
28. Ten Cate O. Competency-based education, entrustable professional activities, and the power of language. J Grad Med Educ 2013;5:6-7.
29. Englander R, Frank JR, Carraccio C, Sherbino J, Ross S, Snell L, *et al*. Toward a shared language for competency-based medical education. Med Teach 2017;39:582-7.
30. Ten Cate O, Hart D, Ankel F, Busari J, Englander R, Glasgow N, *et al*. Entrustment decision making in clinical training. Acad Med 2016;91:191-8.
31. Ten Cate O, Chen HC, Hoff RG, Peters H, Bok H, van der Schaaf M. Curriculum development for the workplace using Entrustable Professional Activities (1): AMEE Guide No. 99. Med Teach 2015;37:983-1002.
32. Carraccio C, Englander R, Gilhooly J, Mink R, Hofkosh D, Barone MA, *et al*. Building a framework of entrustable professional activities, supported by competencies and milestones, to bridge the educational continuum. Acad Med 2017;92:324-30.
33. Ten Cate O. Nuts and bolts of entrustable professional activities. J Grad Med Educ 2013;5:157-8.

Appendix 1: Table of initial assumptions extracted from literature in the assessment system of internal medicine residents at the four inferences of the Kane's validity framework

Validity level	Assumptions	Sub-assumptions
Scoring	The test is properly designed and executed, and also, the scores are a true and accurate representation of the observations. In other words, in addition to the fact that the observations must be done according to the principled and correct methods, the translation of the observations into scores has also been done correctly.	The designers of the various test questions have received the necessary training on the characteristics of each test
		The assessment system has a comprehensive plan and an overall blueprint
		Tests cover different inferences of competence
		Each test has a blueprint and the questions are formulated accordingly
		The minimum passing score in the assessment system under study is determined based on coherent and logical methods and based on scientific principles
		Each of the tests has good internal consistency
		The design of the questions, the holding and execution of each test has proceeded according to scientific principles
Generalization	The tests evaluate appropriate examples of the competencies, expected from the residents, and their results can be generalized to all competencies.	Tests are a good example of the different levels of Miller Pyramid competencies
		Test items are a good example of the content to be evaluated
		The tests have good reliability
		The tests have little error
		The difference between the scores in the tests in general is just due to the real difference between the abilities of the residents and not due to other factors
		Residents' final tests and scores have an acceptable generalizability coefficient
Extrapolation	In addition to being correlated with each other, tests of different inferences of competence also have good predictability for each other	The test results are such that they distinguish the residents of the older years from the residents of the younger years.
		The questions, scenarios and problems of the patients, raised in the tests, correspond to the real world conditions
		There is a good correlation between the scores of the corresponding competencies in different tests
		Low levels of competency scores predict higher levels
Implications	Granting assistants to enter promotion and board exams is consistent with their actual performance throughout the year in the workplace and the results of promotion and board exams.	There is a correlation between the scores obtained in the departmental exams, and the score of the regional or national promotion exam and board exam.
		The scores, obtained in the exams by residents, are correlated with their medical orders
		Test scores show the trend of increasing the experience and ability of residents in each year
		Test scores show the trend of increasing the experience and ability of residents during the course of study

Appendix 2: Proposed methods extracted from literature to evaluate the validity of the system of assessment of specialized residents of internal medicine based on the Kane's framework

Type of test Scoring	Methods and measures required to validate tests at each level of the Kane's framework			
	Generalization	**Extrapolation**	**Implications**	
Written exams				
Objective (pre-progress test)	Checking the training of item designers Checking the implementation of the comprehensive plan of residents' assessments Comparing the questions with blueprint Checking the standard setting for each year Examining the results of test analysis in terms of correlation of questions	Reviewing the results of test analysis Checking the status of sampling questions in Blueprint Checking the generalizability coefficient of tests	Investigating the difference in scores in different years of residency Checking the authenticity of the scenarios Checking the correlation of the corresponding questions in different tests	Checking the correlation of scores with the results of the progress test Assessing the correlation between test/assessment results based on the general opinion of professors
Essay	Examining the quality of questions Checking the standard setting for each year	Checking the sampling of questions Checking the test reliability Checking the generalizability coefficient2 of tests	Checking the correlation of the corresponding questions in different tests	Checking the correlation of scores with the results of the progress test Assessing the correlation between test/assessment results based on the general opinion of professors
OSCE	Checking the training of question designers Checking how to prepare station checklists Completing of the test quality checklist Checking how the stations are arranged for each year of residency Checking how to determine the standard setting of each year	Examining how the curriculum is sampled to determine stations Testing the reliability of the test Investigating the correlation between the scores of the residents of the parallel lines of the test and different times Checking the sources of error in the test Checking the test generalizability coefficient	Investigating the correlation between station scores and corresponding tests Checking the authenticity of the scenarios Checking the difference in grades in different years of residency	Checking the correlation of scores with pre-progress test Checking the correlation of scores with progress test results Assessing the correlation between test/assessment results based on the general opinion of professors
Mini-CEX	Checking the training of question designers Completing the quality checklist of the exam investigating the correlation of assessor's gender and the scores Checking the strictness and lenience of professors Ensuring that the assessor's gender is not related to the score Assessing the satisfaction of residents and evaluators of the test	Frequency of test components (patient type, test setting, disease complexity, test focus type .) Investigating the effective factors in the variation of test scores Checking the test generalizability coefficient	Reviewing the progress of scores in different months Investigating the correlation of corresponding competencies in different tests	Investigating the correlation of scores with the pre- progress test Assessing the correlation between test/assessment results based on the general opinion of professors Reviewing the results of feedbacks
Intra-wards score	Checking the holding according to the comprehensive schedule of residents' exams Checking how to complete the tool	Checking the reliability of scores	Checking the correlation between the corresponding items in different tests	Assessing the correlation between test/assessment results based on the general opinion of professors
Professional behavior score	Checking how to complete the tool	Checking the reliability of scores	Checking the correlation between the corresponding items in different tests	Checking the correlation between the corresponding items in different tests

Contd...

Appendix 2: Contd...

Type of test Scoring	Methods and measures required to validate tests at each level of the Kane's framework			
	Generalization	**Extrapolation**	**Implications**	
Logbook	Checking how the residents complete the logs Checking how the teachers score	Checking the reliability of scores	Checking the correlation between the corresponding items in different tests	Assessing the correlation between test/assessment results based on the general opinion of professors
Record writing score	How to design test tools Checking how to complete the tool	Checking the reliability of scores	Checking the correlation between the corresponding items in different tests	Assessing the correlation between test/assessment results based on the general opinion of professors
Final scores	Checking the conformity of how to calculate the score with the regulations	Investigating the factors of variations in scores Checking the reliability of scores Checking the total generalizability coefficient	Checking the correlation between the scores of different tests	Investigating how minimum pass levels are implemented for each year Evaluating the correlation between the final score and the assessment based on the general opinion of the professors