REVIEW ARTICLE

The Validity Concept in Medical Education: a Bibliometric Analysis

O Conceito de Validade na Educação Médica: uma Análise Bibliométrica

Ruy Guilherme Silveira de Souza[I]
Bianca Jorge Sequeira[I]
Antonio Carlos Sansevero Martins[I]
Angélica Maria Bicudo[II]

KEYWORDS

– Validity.
– Medical Education.
– Assessment.

ABSTRACT

Introduction: *Assessment is a critical part of learning and validity is arguably its most important aspect. However, different views and beliefs led to a fragmented conception of the validity meaning, with an excessive focus on psychometric methods and scores, neglecting the consequences and utility of the test. The last decades witnessed the creation of a significant number of tests to assess different aspects of the medical profession formation, but researchers frequently limit their conclusions to the consistency of their measurements, without any further analysis on the educational and social impacts of the test. The objective of this work is to determine the predominant concept of validity in medical education assessment studies.* **Method:** *The authors conducted a bibliometric research of the literature about studies on the assessment of learning of medical students, to determine the prevalent concept of validity. The research covered a period from January 2001 to august 2019. The studies were classified in two categories based on their approach to validity: (1)" fragmented validity concept" and (2)" unified validity concept". To help with validity arguments, the studies were also classified based on Miller's framework for clinical assessment.* **Results:** *From an initial search resulting in 2823 studies, 716 studies were selected based on the eligibility criteria, and from the selected list, of which 693 (96,7%) were considered studies of the fragmented validity concept, which prioritized score results over an analysis of the test's utility, and only 23 studies (3,2%) were aligned with a unified view of validity, showing an explicit analysis of the consequences and utility of the test. Although the last decade witnessed a significant increase in the number of assessment studies, this increase was not followed by a significant change in the validity concept.* **Conclusions:** *This bibliometric analysis demonstrated that assessment studies in medical education still have a fragmented concept of validity, restricted to psychometric methods and scores. The vast majority of studies are not committed to the analysis about the utility and educational impact of an assessment policy. This restrictive view can lead to the waste of valuable time and resources related to assessment methods without significant educational consequences.*

[I]Universidade Federal de Roraima, Boa Vista, Roraima, Brazil.
[II]Universidade Estadual de Campinas, Campinas, São Paulo, Brazil.

DOI: https://doi.org/10.1590/1981-5271v44.4-20200055.ING

PALAVRAS-CHAVE

– Validade.

– Educação Médica.

– Avaliação.

RESUMO

Introdução: *Avaliação é uma parte crítica da aprendizagem, e validade é sem dúvida seu aspecto mais importante. No entanto, diferentes visões e crenças levaram a uma concepção fragmentada do significado de validade, com um foco excessivo nos métodos psicométricos e escores, negligenciando a utilidade do teste. As últimas décadas testemunharam a criação de um número significativo de testes para avaliar diferentes aspectos da formação da profissão médica, mas os pesquisadores frequentemente limitam suas conclusões à consistência de suas medidas, sem nenhuma análise adicional sobre os impactos educacionais e sociais do teste. O objetivo deste trabalho é determinar o conceito predominante de validade nos estudos de avaliação em educação médica.* ***Método:*** *Foi realizada uma pesquisa bibliométrica da literatura de estudos sobre avaliação da aprendizagem de estudantes de Medicina para determinar o conceito prevalente de validade. A pesquisa abrangeu o período de janeiro de 2001 a agosto de 2019. Os estudos foram classificados em duas categorias: 1. "conceito de validade fragmentada" e 2. "conceito de validade unificada". Para ajudar nos argumentos de validade, os estudos também foram classificados com base na estrutura de Miller para avaliação clínica.* ***Resultados:*** *A partir de uma pesquisa inicial que resultou em 2.823 estudos, selecionaram-se 716 com base nos critérios de elegibilidade, e consideraram-se 693 (96,7%) estudos com conceito fragmentado de validade que priorizavam os resultados dos escores em detrimento de uma análise da utilidade do teste, e apenas 23 (3,2%) foram alinhados com uma visão unificada de validade, apresentando uma análise explícita das consequências e da utilidade do teste. Embora a última década tenha testemunhado um aumento expressivo de estudos sobre avaliação, esse crescimento não foi acompanhado por uma mudança significativa do conceito de validade.* ***Conclusões:*** *Esta análise bibliométrica demonstrou que os estudos sobre avaliação de aprendizagem em educação médica têm um conceito fragmentado de validade, limitados aos métodos psicométricos e escores. A grande maioria dos trabalhos não está comprometida com uma análise sobre a utilidade e o impacto educacional de uma política de avaliação. Essa visão restritiva pode levar à perda de tempo e recursos valiosos com métodos de avaliação sem consequências educacionais significativas.*

Received on 3/5/20

Accepted on 9/26/20

INTRODUCTION

Validity is arguably the most important aspect of any kind of assessment[1-4], however the overwhelming number of different concepts and beliefs about it has led to some confusion regarding its significance[5]. Traditionally, in philosophy, the term is derived from the Latin word "Validus" (meaning "strong" or "worth") and is a fundamental aspect of logic used to provide deductive arguments about a fact or idea. At the transition to the XX century, it emerged in the field of education and psychology as a strategy to justify a growing number of tests in the form of structured assessment, and since they were used to support complex and important decisions, from selection process to educational policies, the quest for validity unleashed a movement of tremendous empirical effort to demonstrate the intrinsic efficacy of these tests. The introduction of the "correlation coefficient" by Karl Pearson in 1896, permitted the estimation of the correlation between different criteria, and not long after that, different psychometric strategies started to be used as validity arguments[6]. Soon, different views and constructs resulted in the fragmentation of the validity concept with an increasing number of "types of validity"[7], and in 1955 the "Joint Committee of American Psychological Association (APA)" was officially using four distinctive varieties: "content validity", "predictive validity", "concurrent validity" and "construct validity". This fragmented approach overestimated the process of collecting evidence through measurements, neglecting the analysis of the consequences and use of the test. Cronbach and Meehls revised these different categories and created what would become the cornerstone of the current validity concept[8]. They proposed that the validation process should not be limited to evidence gathering, but demanded an extensive analysis of these findings based on an explicit statement of the proposed interpretation, or in Kane's words "*the variable of interest is not out there to be estimated; the variable of interest has to be defined or explicated*"[9]. Since this concept could be applied to any kind of construct and validation process, it paved the way to a unified view of validity, where "*all validity is construct validity*"[1,10].

Messick, who represented one of the spearheads of the unification movement, proposed that validity should be guided by two questions: the first one, of scientific nature, should be concerned with the psychometrics properties of the test, and the second and essential one, of ethical nature, could only be answered by an extensive analysis of the potential utility and consequences of a test in "*terms of human values*"[11]. Cronbach summarized this new idea stating that "*One validates, not a test, but an interpretation of outcomes from measurement procedure*"[12].

The last decades witnessed a significant number of tests created to assess different aspects of the medical profession formation (cognitive, skills, attitude)[13,14]. In this scenario, where medical teachers are hard-pressed to attest the efficacy of their assessment methods, medical education risks going back to an era characterized in Cronbach's words as '*sheer exploratory empiricism*"[3], where measurement and the consistency of its results are more important than the assessment itself, and which will result in the progressive reduction of the validity meaning, from a philosophical exercise to an attribute of a measure.

The consequence of such a restrictive view is the excessive use of different type of tests, driven not by an educational purpose but by the impulse to follow the latest fad, without much reflection on their utility

and educational impacts[15].

The objective of this analysis is to determine the predominant concept of validity used in medical education assessment studies. For this purpose, a bibliometric research was conducted in the literature, covering the last two decades of published assessment works to construct a metric with quantitative indicators on the utility of assessment studies and the perceived change in the concept of validity along the period.

METHODS

Search strategy

The data for this bibliometric analysis was collected based on a search for articles indexed in the PubMed database on assessment of knowledge in medical education, covering a period from January 2001 to august 2019, using the terms "validity" and & or "medical education" and & or "assessment". A review group was created to ensure expertise in medical education and research methodology, consisting of two clinicians with postgrad formation on medical education and a university researcher. The members of the review group independently applied the inclusion and exclusion criteria (described below) to make up a combined list. A preliminary analysis of the literature was performed to determine the existence of evidence in similar searches and reviews, by searching the databases of Cochrane and BEME reviews.

Inclusion and exclusion criteria

The inclusion criteria were as follows: (1) study design: all studies on learning assessment that included the validation process in its methodology; (2) population: medical students at undergraduate and graduate levels; (3) educational intervention: studies on learning assessment in the cognitive, skills and attitudinal domains. The exclusion criteria were as follows: (1) study design: systematic reviews and reviews and studies published before 2001; (2) population: studies that did not focus exclusively on medical students.

Study selection and classification

A two-stage process was employed for selection and classification. Initially, based on the first broad search based on the eligibility criteria, the authors made a secondary selection, where studies of which validation process was not explicit in their methods section were excluded.

Subsequently, the selected studies were then classified in two categories and three levels, based on their approach to validity. Category one represents studies with the fragmented validity concept, and was subdivided in two levels: "level one", where the interpretation of validity was restricted to the reporting of test scores, and "level two" where test scores were followed by some kind of inference but without an explicit statement about the utility and consequences of the test. Category two represents studies with a unified validity concept and is entirely constituted by 'level three", where the utility and consequences of the test are made explicit (Table 1).

The authors made a preliminary selection and categorization of studies independently, followed by a consensus process to determine the final classification. Any disagreement on the classification were resolved by subsequent discussion in the panel until a reliable consensus was reached. Furthermore, 20% of the selected studies were randomly submitted to double evaluation for quality assurance and prevention of bias.

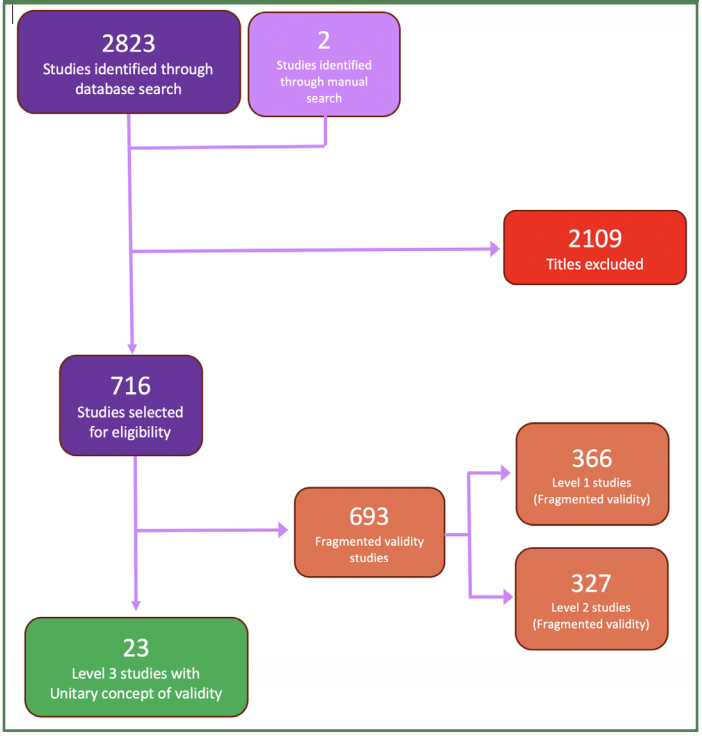
The authors also made a manual search, and based on the article "The top-cited articles in medical education: A bibliometric analysis"[18], a list was created with the three most frequently cited journals, to serve as a basis for the manual research. Figure 1 represents the research flowchart, from the broad search at the PUBMED database to the final classification of the studies.

After the final selection and classification, the studies were also classified based on the Miller framework for clinical assessment[13], to serve as an accessory for validity arguments. For this purpose, studies were sorted into two main groups: (1) "written, oral or computer-based assessment" for the "Knows-Knows how" first two steps of the Miller's pyramid, and (2) "performance or hands-on assessment" for the "Shows how-Does" top two steps.

Table 1
Classification of studies

	Classification of studies according to the validity concept	
Categories	**Levels**	**Criteria**
1-Studies with the fragmented validity concept	Level one	-Studies that do not make explicit the utility of the test being limited to the reporting of results and scores.
	Level two	-Studies that do not make explicit the utility of the test, being limited to the reporting of results and scores, with some kind of inference.
2-Studies with unitary validity concept	Level three	-Studies that use the results and inferences of the validation process to make useful pedagogical decisions for the school.

Figure 1
Research flowchart



DOI: https://doi.org/10.1590/1981-5271v44.4-20200055.ING

Statistical correlation between kinds of assessment in the two decades was assessed by the chi-square test. Statistical analysis was performed using the freeware R 3.2.0.

RESULTS

From an initial broad search resulting in 2823 studies, 716 studies were selected based on the eligibility criteria. Of these, 693 (96.7%) were considered studies with the fragmented concept of validity. A total of 366 studies (50,83%) were classified as "level one", which limited their validity analysis to the resulting score of the validation process without any kind of inference or statement about the consequences or utility of the test. In 327 studies (45,94%), the results were accompanied by some kind of inference about the results but without any report or discussion about the utility of the test and were classified as 'level two"; and only 23 studies (3.2%) met the criteria for "level three", where the authors presented the results of the validation process aligned with explicit analysis of the consequences and utility of the test. Figure 2 shows the temporal distribution of selected studies grouped according to the three levels of validity concept.

The temporal distribution showed a significant increase in the number of validity studies in the last decade of the present century. The number of studies increased from 179 studies in "decade ONE/XXI", to 537 studies in "decade TWO/XXI" (P<0,001). This significant increase in the number of validity studies from one decade to another, was not accompanied by a proportional change in the level of validity. Although there was also an almost two-fold increase in the proportion of level 3 studies, from 1.78% in the first decade to 3.31% in the second decade, this increase was not significant (P=0.356). Thus, 96,27% studies in decade TWO/XXI were still considered to have a fragmented validity concept (Table 2).

Based on Miller framework for clinical assessment[13], 415 studies were classified as 'written, oral or computer-based assessment" and 301 studies were classified as "performance or hands-on assessment". In a temporal distribution, the decade "ONE/XXI" had a total of 179 works, with 104 (58.11%) classified as "written, oral or computer-based assessment", and 75 (41.89%) as "performance or hands-on tests". Of the 537 works of the decade "TWO/XXI", 311 (57.91%) were classified as "written, oral or computer-based assessment" and 226 (42.09%) studies were classified as

Figure 2
Temporal distribution of studies

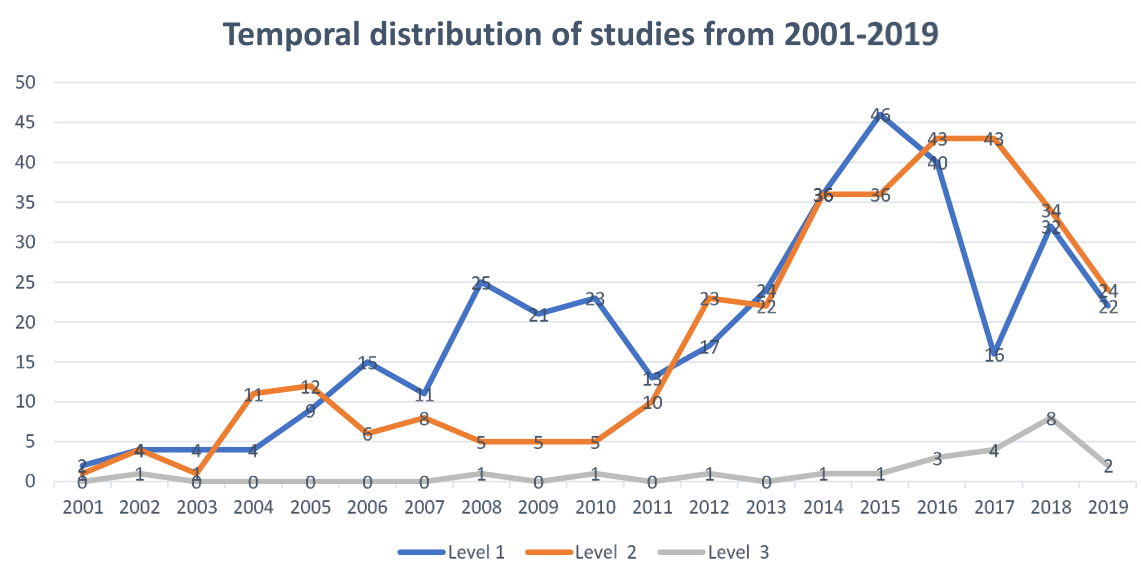


Table 2
Validity studies in the first two decades of the 21st century

Decades	Decade I -XXI (2001-2010)			Decade II -XXI (2011-2019)			
Articles	N	Percentage Mean	SD	N	Percentage Mean	SD	Sig*
Level 1	118	63.07	13.82	248	46.2	7.19	0.03
Level 2	58	35.15	13.67	269	50.73	5.95	0.04
Category 1 (L1 + L2)	176	98.22	2.54	517	96.94	2.69	0.444
Category 2(Level 3)	3	1.78	2.54	20	3.31	2.67	0.356
Mean	17.9		6.98	59.6		16.41	0.001

*independent samples t-test (CI=95%)

Table 3
Distribution of studies based on Miller framework for clinical assessment

	Decade I -XXI		Decade II -XXI		
Articles	N	%	N	%	Sig*
Written/Oral/Computer-based	104	58.11	311	57.91	p=1
Performance/Hands-on	75	41.89	226	42.09	
Total	179	100.00	537	100.00	

*chi-square(CI=95%)

"performance or hands-on tests". Although there was a significant increase in number of studies in the last decade, there was no significant change in the proportion between "written, oral or computer-based assessment" studies and "performance or hands-on assessment" (Table 3).

DISCUSSION

There has been an unquestionable surge of assessment research in medical education, but this increase in the number of studies does not necessarily correlate with an equivalent educational impact. In the same way that adult learning captivated the attention of medical educators in the second half of the 20th century, we have arrived at the 21st century with assessment papers representing almost half of the most cited articles[18], and with the increasing interest in the latest trends in medical education, such as outcome-based curriculum[19] and assessment of entrustable professional activities[20], where assessment has a prominent role, we can only expect a continuous increase. In contrast with this growing popularity, is the scarcity of validity research with a unitary unified view, which recognizes the consequences of a test policy. This search demonstrated that a fragmented view of validity is still prevalent in most study designs. Almost the totality of works (96.7%) presented a fragmented concept of validity and half of all articles (50.83%) were limited to the isolated demonstration of results from the validation process, leaving aside the most important aspect of validity, which is the score interpretation and its subsequent use[1]. It is also noteworthy that the impressive escalation in the number of assessment studies, documented in this work, was not followed by an equivalent increase in the number of studies with a unitary unified view of validity, demonstrating a renewed interest on the assessment but with the same old fragmented view of validity.

The biological and quantitative heritage in biomedical sciences has a strong influence on medical education, contributing to a mechanistic view of assessment, frequently described as a tool or instrument; however, assessment should be viewed in a much broader sense than measurement. Royal conducted a general Pub Med search in a five-year period, and found the term "reliable instrument" over two thousand times[21]. This mechanistic approach is supported by the misconception that learning is linear and independent of context, and that it could always be objectively estimated by reliable instruments used by trained assessors[22], but reliability, although necessary, is never sufficient for a valid argument supporting an educational decision[23]. The medical education literature seems to approach validity as some kind of property of a test, when in fact, validity is about the meaning of test scores and, especially, of its use[4,11,24].

In the same way that in the mid-1990s Bloom pointed an excessive focus on curriculum reform without paying attention to the "learning environment", leading to an era of "*reform without change*"[29], the excessive experimentation on different assessment methods without giving much thought to the educational consequences may lead to discredit and obsolescence of many instruments, inaugurating an era of "*useless tests*"[24]. The pervasive view of assessment as a box of different and disconnected tools, and the inability to differentiate the validation process of these instrument from the real meaning of validity, make medical educators forget that a medical school is a social environment, with all the peculiarities that can make the most reliable test have a very poor reception if one does not observe the local idiosyncrasies. One simply cannot make a judgement based on a single measurement, and medical schools must develop a more comprehensive view of assessment, not as a tool, but as a program effectively integrated to the educational program[14].

"From prediction to explanation"

In a significative number of studies selected in this search, it is clear that many researchers are satisfied solely by the consistency of their test scores, without any further analysis, and it is easy to understand the appeal of reliability over validity in a context where the scientific aspects of medical education casts a long shadow over the humanistic values[25]. Even Cronbach, one of validity champions, in one of his many important contributions, made a significant extension of the reliability theory to the generalizability theory[26], and in the same way for medical teachers with a strong biomedical formation, generalization based solely on the consistency of scores stimulates the replication of experiments without much reflection on current developments on the learning theory and its ethical and social impacts[27].

Early exposure of learners to practice[28], the growing importance of work-based assessment[29] and the complexity of health care all point to the limitations of a statistical-based approach to assess learning in the medical profession, pointing to the dawn of a post-psychometric age[30]. It is necessary to have a "*shift from numbers to words*"[22] and it is time for medical teachers to look for the intrinsic value of a test, based not only on the consistency of its scores, but also on the utility and consequences of its use. In this sense, educators recognize that one cannot assess learning based solely on test scores,[31,32] and the qualitative assessment can offer solutions to many limitations of numerical values[33,34], contributing to the validity argument. This bibliometric analysis demonstrated that the growing interest in performance assessment has not been followed by an equivalent increase in the number of studies, and with the increasing number of medical schools investing in skills labs and high fidelity simulation[35], it is urgent to expand the view of validity, to a qualitative and ecological dimension, where the observed behavior in the laboratory can be generalized to the workplace[36,37].

Slowly, medical education literature is beginning to shift its attention to a unified view of validity[38,39], but the prevalent understanding is still fragmented, with validity frequently confused with the validation process, prioritizing empirical analysis and scores to the detriment of a social dimension of a test[40,41].

CONCLUSION

Validity is "the most important criterion" in a test[42], but it is frequently underestimated and compared to subordinate criteria. This systematic

search demonstrated that assessment studies in medical education are still far from a unified view of validity, and not committed to an extensive analysis involving all possible consequences of a test policy, especially those related to social and educational impacts. This restrictive view can lead to the waste of valuable time and resources in assessment methods, without any significant educational consequence. Future studies should prioritize assessment research specifically tailored to the needs of the school and integrated to the educational program, shifting the focus from a culture of replication of assessment methods to a social evaluation, where aspects like utility and educational impact should be the primary goals.

REFERENCES

1. Downing SM. Validity: on the meaninful interpretation of assessment data. Med Educ. 2003;37(9):830-7. doi: 10.1046/j.1365-2923.2003.01594.x.
2. Downing SM. Validity threats: overcoming interference with proposed interpretation of assessment data. Medic Educ. 2004;38(3):327-33. doi: 10.1046/j.1365-2923.2004.01777.x.
3. Newton PE, Shaw SD. Validity in Educational & Psychological Assessment. Fisrt ed. Cambridge: Cambridge Assessment; 2014.
4. Messick S. Validity of Psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. New Jersey: Educational Testing Service; 1994.
5. Bergmann AC, Childs RA. When I say... validity Argument. Med Educ. 2018;52(10):1003-4. doi: 10.1111/medu.13592.
6. Ratner B. The correlaciona coefficient: its values range between +1/-1, or do they? Journal of Targeting, Measurement and Analysis for Marketing. 2009;17:139-142. doi: 10.1057/jt.2009.5.
7. Souza AC, Alexandre NMC, Guirardello EB. Propriedades psicométricas na avaliação de instrumentos: avaliação da confiabilidade e validade. Epidemiol Serv Saude. 2017;26(3):649-59.
8. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychological Bulletin. 1955;52(4):281-302.
9. Kane M. Currents concerns in validity theory. Annual Meeting of the American Educational Research Association-AERA. New Orleans; 2000.
10. Messick S. Validity. In: Linen R, editor. Educational Measurement. 3rd ed. New York: American Council on Education and Macmillan; 1989. p.13-104.
11. Anderson S, Messick S. Social Competency in Young Children. Dev Psychol. 1974;10(2):282-93.
12. Cronbach LJ. Test validation. In: Thorndike RL, ed. Educational Measurement. 2nd edition ed. Washington DC: American Council on Education; 1971. p. 443-507.
13. Miller GE. The assessment of clinical skills/ competence/ performance. Acad Med. 1990(suppl):S63-S67.
14. van der Vleuten CPM. Revisiting 'Assessing professional competence: from methods to programmes'. Med Educ. 2016;50(9):885-8. doi: 10.1111/medu.12632.
15. Norman G. Editorial-Inverting the pyramid. Adv in Health Sci Educ. 2005;10:85-8.
16. Moher D, Liberati A, Tetzlaff J, Altman D. Preferred Reporting Items for Systematic Review and Meta-Analysis: the Prisma Statement. BMJ. 2009;6(7):e1000097. doi:10.1371/journal.pmed.1000097.
17. Cook D, West C. Conducting systematic reviews in medical education: a stepwise approach. Med Educ. 2012;46(10):943-52.
18. Azer S. The Top-Cited Articles in Medical education: A Bibliometric Analysis. Acad Med. 2015;90(8):1-9.
19. Harden RM. Outcome-based Education: the future is today. Med Teach. 2007;29(7):625-9. doi: 10.1080/01421590701729930.
20. Ten Cate O. Entrustability of professional activities and competency-based training. Med Educ. 2005;39(12):1176-7.
21. Royal KD. Four tenets of modern validity theory for medical education assessment and evaluation. Adv Med Educ Pract. 2017;8:567-9. doi: 10.2147/AMEP.S139492.
22. Govaerts M, Van der Vleuten C. Validity in work-based assessment: expanding our horizons. Med Educ. 2013;47(12):1164-74. doi: 10.1111/medu.12289.
23. Cook D, Beckman T. Current Concepts in validity and reliability for psychometrics Instruments: Theory and Applications. Am J Med. 2006;119(2):166.e7-16. doi: 10.1016/j.amjmed.2005.10.036.
24. Sireci SG. On the validity of useless tests. Assessment in Education: Principles, Policy & Practices. 2016;23(2):226-35. doi: 10.1080/0969594X.2015.1072084.
25. Cook M, Irby DM, Sullivan W, Ludmerer KM. American Medical Education 100 Years after the Flexner report. N Engl J Med. 2006;355(13):1339-44.
26. Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. The dependability of behavioral measurements: theory of generalizability for scores and profiles. New York: Wiley; 1972.
27. Bleakly A. Broadening conceptions of learning in medical education: the message from teamworking. Med Educ. 2006;40(2):150-7. doi: 10.1111/j.1365-2929.2005.02371.x.
28. Souza R, Sansevero A. Introducing early clinical experience in the curriculum. In: Bin Abdulrahman K, ed. Routledge international handbook of Medical Education. New York: Routledge; 2016.
29. Norcini J. Work based assessment. BMJ. 2003;326(7392):753-5. doi: 10.1136/bmj.326.7392.753.
30. Hodghes B. Assessment in the post-psychometric era: learning to love the subjective and collective. Med Teach. 2013;35(7):564-8.
31. Ginsburg S, McIlroy J, Oulanova O, Eva K, Regehr G. Toward authentic clinical evaluation: Pitfalls in the pursuit of competency. Acad Med. 2010;85(5):780-6. doi: 10.1097/ACM.0b013e3181d73fb6.
32. Shuwirth L, van der Vleuten CPM. A plea for new psychometric models in educational assessment. Med Educ. 2006;40(4):296-300. doi: 10.1111/j.1365-2929.2006.02405.x.
33. Cook DA, Kuper A, Hatala R, Ginsburg SA. When Assessment Data Are Words: Validity Evidence for Qualitative Educational Assessments. Acad Med. 2016;91(10):1359-69.
34. Mann KV. Theoretical perspectives in Medical education: past experiences and future possibilites. Med Educ. 2011;45(1):60-8. doi: 10.1111/j.1365-2923.2010.03757.x.
35. Massoth C, Röder H, Ohlenburg H, Hessler M, Zarbock A, Pöpping DM, et al. High-fidelity is not superior to low-fidelity simulation but leads to overconfidence in medical students. BMC Medical Education. 2019;19(1):29. doi: 10.1186/s12909-019-1464-7.
36. Schmuckler MA. What is ecological validity? A dimensional Analysis. Infancy. 2001;2(4):419-36.
37. Sturman MC, Cheramie RA, Cashen LH. The impact of job complexity

and performance measurement on the temporal consistency, stability, and test-retest reliability of employee job performance ratings. J Appl Psychol. 2005;90(2):269-83.

38. 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 in Health Sci Educ. 2013;19(2):233-50. doi: 10.1007/s10459-013-9458-4.
39. Pugh DM, Wood TJ, Boulet JR. Assessing Procedural Competence: validity considerations. Simulation in health care. 2015;10(5):288-94.
40. Gulliksen H. Intrinsic validity. American Psychologist . 1950;5(10):511-7.
41. Marceau M, Gallagher F, Young M, St-Onge C. Validity as a social imperative for assessment in health professions education: a concept analysis. Med Educ. 2018;52(6):641-53. doi: 10.1111/medu.13574.
42. Koretz D. Measuring up: What Educational Testing Really tell Us. Cambridge MA: Harvard University Press; 2006.

AUTHORS' CONTRIBUTION

Ruy Guilherme Silveira de Souza, Bianca Jorge Sequeira and Antonio Carlos Sansevero Martins participated in the research, writing and review of the manuscript. Angélica Maria Bicudo participated in the review.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

ADDRESS FOR CORRESPONDENCE

Ruy Guilherme Silveira de Souza. Rua itaúba,1173, Caçari, Boa Vista, RR, Brasil. CEP: 69307-610.

E-mail: ruysouza28@gmail.com