Reliability across content areas in progress tests assessing medical knowledge: a Brazilian cross-sectional study with implications for medical education assessments

Pedro Tadao Hamamoto Filho[I], Miriam Hashimoto[II], Alba Regina de Abreu Lima[III], Leandro Arthur Diehl[IV], Neide Tomimura Costa[V], Patrícia Moretti Rehder[VI], Samira Yarak[VII], Maria Cristina de Andrade[VIII], Maria de Lourdes Marmorato Botta Hafner[IX], Zilda Maria Tosta Ribeiro[X], Júlio César Moriguti[XI], Angélica Maria Bicudo[XII]

Universidade Estadual Paulista, Faculdade de Medicina de Botucatu, São Paulo, Brazil

[I]PhD. Physician, Assistant Professor, Departament of Neurology, Psychology and Psychiatry, Faculdade de Medicina de Botucatu, Universidade Estadual Paulista (UNESP), Botucatu (SP), Brazil.
ⓘ https://orcid.org/0000-0001-6436-9307

[II]PhD. Physician, Assistant Professor, Departament of Pediatrics, Faculdade de Medicina de Botucatu, Universidade Estadual Paulista (UNESP), Botucatu (SP), Brazil.
ⓘ https://orcid.org/0000-0001-6969-9161

[III]PhD. Biologist, Assistant Professor, Departament of Molecular Biology, Faculdade de Medicina de São José do Rio Preto (FAMERP), São José do Rio Preto (SP), Brazil.
ⓘ https://orcid.org/0000-0003-4332-4059

[IV]PhD. Physician, Assistant Professor, Departament of Internal Medicine, Universidade Estadual de Londrina (UEL), Londrina (PR), Brazil.
ⓘ https://orcid.org/0000-0002-9958-7213

[V]PhD. Physician, Assistant Professor, Departament of Internal Medicine, Universidade Estadual de Londrina (UEL), Londrina (PR), Brazil.
ⓘ https://orcid.org/0000-0001-6624-2043

[VI]PhD. Physician, Assistant Professor, Departament of Tocogynecology, Universidade Estadual de Campinas (UNICAMP), Campinas (SP), Brazil.
ⓘ https://orcid.org/0000-0003-4530-1109

[VII]PhD. Physician, Assistant Professor, Departament of Dermatology, Universidade Federal de São Paulo (UNIFESP), São Paulo (SP), Brazil.
ⓘ https://orcid.org/0000-0002-5657-6645

[VIII]PhD. Physician, Assistant Professor, Departament of Pediatrics, Universidade Federal de São Paulo (UNIFESP), São Paulo (SP), Brazil.
ⓘ https://orcid.org/0000-0003-4519-0613

[IX]PhD. Physician, Assistant Professor, Academic Assessment Center, Faculdade de Medicina de Marília (FAMEMA), Marília (SP), Brazil.
ⓘ https://orcid.org/0000-0002-5471-4434

[X]PhD. Physician, Assistant Professor, Academic Assessment Center, Faculdade de Medicina de Marília (FAMEMA), Marília (SP), Brazil.
ⓘ https://orcid.org/0000-0001-7368-3154

[XI]PhD. Physician, Associate Professor, Departament of Internal Medicine, Universidade de São Paulo (USP), Ribeirão Preto (SP), Brazil.
ⓘ https://orcid.org/0000-0001-5499-3552

[XII]PhD. Physician, Full Professor. Departament of Pediatrics, Universidade Estadual de Campinas (UNICAMP), Campinas (SP), Brazil.
ⓘ https://orcid.org/0000-0003-3043-5147

ABSTRACT

BACKGROUND: Brazilian medical schools equitably divide their medical education assessments into five content areas: internal medicine, surgery, pediatrics, obstetrics and gynecology, and public health. However, this division does not follow international patterns and may threaten the examinations' reliability and validity.

OBJECTIVE: To assess the reliability indices of the content areas of serial, cross-institutional progress test examinations.

DESIGN AND SETTINGS: This was an analytical, observational, and cross-sectional study conducted at nine public medical schools (mainly from the state of São Paulo) with progress test examinations conducted between 2017 and 2023.

METHODS: The examinations covered the areas of basic sciences, internal medicine, surgery, pediatrics, obstetrics and gynecology, and public health. We calculated reliability indices using Cronbach's α, which indicates the internal consistency of a test. We used simple linear regressions to analyze temporal trends.

RESULTS: The results showed that the Cronbach's α for basic sciences and internal medicine presented lower values, whereas gynecology, obstetrics, and public health presented higher values. After changes in the number of items and the exclusion of basic sciences as a separate content area, internal medicine ranked highest in 2023. Individually, all content areas except pediatrics remained stable over time.

CONCLUSIONS: Maintaining an equitable division in assessment content may lead to suboptimal results in terms of assessment reliability, especially for internal medicine. Therefore, content sampling of medical knowledge for general assessments should be reappraised.

INTRODUCTION

Traditionally, Brazilian medical schools have divided general medical education assessments into five content areas: internal medicine, surgery, pediatrics, obstetrics and gynecology, and public health. This division was the consequence of a resolution passed in 2000 by the National Medical Residency Commission,[1] which defined the organization of the selection processes nationally. Despite the several modifications to the resolution in subsequent years, the equitable division of the number of items between the five content areas has remains untouched.

Therefore, all medical selection processes for general specialties (e.g., internal medicine, pediatrics, and general surgery) and direct-access specialties (e.g., anesthesiology, ophthalmology, and radiology) are legally obligated to use this division. Similarly, many undergraduate medical curricula follow this division, either for conducting assessments, such as inter-institutional progress testing,[2] or for organizing clerkship rotations.[3]

However, this equitable division of item numbers does not conform to international patterns. In the Netherlands, progress testing uses a two-dimensional blueprint that includes elements from disciplines (e.g., surgery, dermatology, pediatrics, and physiology) and categories (e.g., the respiratory and musculoskeletal systems).[4] The German Progress Test Medizin is blueprinted according to organ systems with different percentages for each (e.g., 11% for the cardiac system and 4.5% for skin).[5] Similarly, Step 2 of the United States Medical Licensing Examination includes differential weighing of content areas (e.g., 8–10% for the cardiovascular system and 4–6% for pregnancy, childbirth, and puerperium).[6]

Furthermore, the numbers of items across different content areas have implications for examination reliability, which is the reproducibility of assessment outcomes over time or on specific occasions, that is, the consistency of measurements. Therefore, owing to a large component of random errors, the data resulting from low-reliability assessments may threaten the generalization and interpretation of the results.[7]

OBJECTIVE

This study aimed to assess the reliability indices of the content areas of serial, cross-institutional progress test examinations conducted in the state of São Paulo between 2017 and 2023.

METHODS

Study design

We conducted an analytical, observational, cross-sectional study based on data from a retrospective database of inter-institutional progress test examinations held between 2017 and 2023 in nine public medical schools (mainly from the state of São Paulo). We considered only the grades achieved by sixth-year medical students as the test was designed at the level of a recently graduated physician. **Table 1** presents the number of participating students for each year.

Ethical considerations

We used only secondary data from the examinations and did not identify individual students; therefore, approval from an ethical review board was not necessary according to national legislation.

Settings and progress test information

The participating schools joining a consortium were Universidade Estadual Paulista (UNESP), Universidade Estadual de Campinas (UNICAMP), Universidade de São Paulo (USP – Ribeirão Preto), USP – Bauru, Universidade Federal de São Paulo (UNIFESP), Universidade Federal de São Carlos (UFSCAR), Faculdade de Medicina de Marília (FAMEMA), Faculdade de Medicina de São José do Rio Preto (FAMERP), and Universidade Estadual de

Table 1. Number of participating sixth-year medical students across the observational period

Year	Number
2017	687
2018	742
2019	713
2020	712
2021	666
2022	726
2023	525
Total	4771

Londrina (UEL). Universidade Regional de Blumenau (FURB) participated in the progress test examinations from 2017 to 2022. Further, Universidade de São Paulo, Bauru campus was inaugurated in 2018; consequently, their sixth-year students participated in the progress test examinations only in 2023.

From 2017 to 2022, the progress test examination consisted of 120 multiple-choice questions equally divided into 20 items from basic sciences, internal medicine, pediatrics, surgery, gynecology and obstetrics, and public health. The use of basic sciences as a content area was related to the consortium's specific history. However, in 2023, basic sciences-related content was distributed among the other five areas, and the number of items was changed to emphasize internal medicine more. Therefore, in 2023, the items were 34 for internal medicine, 23 for pediatrics, 23 for surgery, 20 for gynecology and obstetrics, and 20 for public health. The items were based on clinical vignettes and focused on applied knowledge rather than on the retrieval of memorized information.

Data analysis

We calculated reliability indices using Cronbach's α coefficients,[8] which provide a measure of the internal consistency of a scale or test and range from 0 (low consistency) to 1 (high consistency). Tavakol and Dennick have stated that "*internal consistency describes the extent to which all the items in a test measure the same concept or construct and hence it is connected to the interrelatedness of the items within the test.*"[9] We calculated the α coefficients of each content area for each examination year using the following formula:

$$\alpha = \left(\frac{k}{k-1}\right) x \left(1 - \frac{\sum_{i=1}^{k} s_i^2}{s_{sum}^2}\right)$$

where k is the number of items in the test; s_i^2 is the variance of each item, and s_{sum}^2 is the variance in total scores for each respondent.

We tested the differences between the mean α coefficients using one-way analysis of variance (ANOVA) and calculated temporal trends in α coefficients using a simple linear regression model. The statistical significance level was set at P = 0.05. We performed statistical analyses using SPSS, version 24.0 (IBM Corp., Armonk, New York, United States) and Prism 9 for MacOS (version 9.5.0, GraphPad Software, San Diego, California, United States).

RESULTS

In terms of the absolute values of the α coefficients, basic sciences and internal medicine presented lower values, whereas gynecology and obstetrics, and public health presented higher values (**Figure 1**). After ranking the content areas according to a
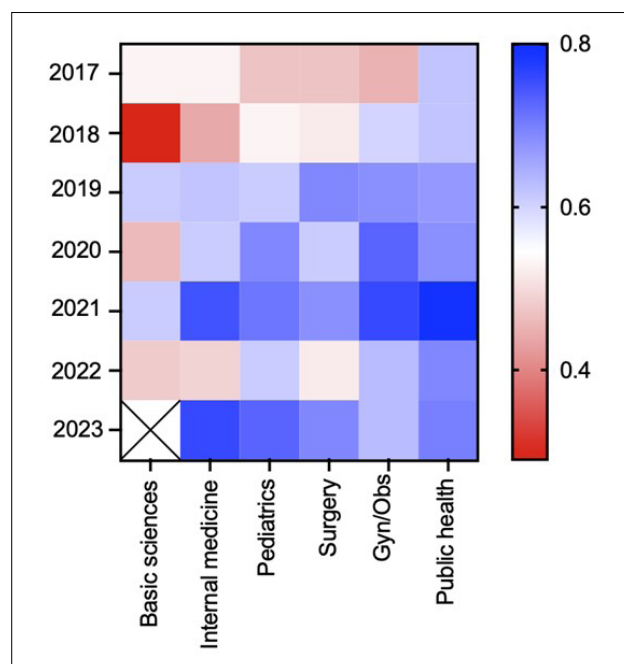
Figure 1. Heat map of alpha coefficient values for the six content areas between 2017 and 2023.

top-down classification of the α coefficients, we found that basic sciences ranked last in almost all years. Internal medicine had intermediate to low positions. Further, pediatrics and surgery had intermediate positions, while public health and gynecology and obstetrics were the highest-ranked areas. After the changes in the number of items and the absence of basic sciences as a separate content area in 2023, internal medicine had the highest position.

The mean α coefficients from 2017 to 2023 were 0.497, 0.600, 0.621, 0.597, 0.640, and 0.683 for basic sciences, internal medicine, pediatrics, surgery, gynecology and obstetrics, and public health, respectively. We found no statistically significant difference in the mean values ($F = 2.432$, $P = 0.054$).

Figure 2 shows the analysis of the temporal trends of each area, which revealed a significant difference in the intercept values of the lines, with basic sciences and internal medicine presenting the lowest values ($F = 2.897$, $P = 0.027$). That is, the initial values of the trend lines were significantly lower in these two content areas. Individually, all content areas except pediatrics remained stable over time, and pediatrics presented a significant upward trend despite a slight decrease in 2022.

Regarding the slopes, that is, the annual change in trend line characteristics, we found no difference between the content areas ($F = 0.206$, $P = 0.957$). Therefore, despite the differences in the initial values, the lines remained parallel to each other, with a pooled slope of 0.026, which indicates a slight increase in the examinations' overall consistency (**Table 2**).

DISCUSSION

The use of progress testing by medical schools is increasing in Brazil as feedback for students, faculty, and institutions effectively improves medical education.[2,10,11]

Created in the Netherlands and USA in the 1970s, progress testing began to be used in Brazil at Universidade de São Paulo and Universidade Estadual de Londrina between the late 1990s and early 2000s. The first progress testing examinations at Universidade de São Paulo consisted of 100–130 items divided into three areas: basic sciences, clinical sciences, and clerkship rotations. This division followed the traditional organization of medical curricula, and the number of items in each area was weighted according to the disciplines' workload.[11] At Universidade Estadual de Londrina, the test was divided into six areas: basic sciences, internal medicine, pediatrics, surgery, gynecology and obstetrics, and public health.[13]

This organization of curricula and assessments has been used in Brazil since the early 1990s owing to the immense efforts of the Interinstitutional Commission for the Evaluation of Medical Education (CINAEM).[14] This commission has played an important role in restructuring medical education in Brazil, especially in assessing medical school processes.[15] Their work culminated in the publication of the National Curricular Guidelines in 2001,[16] which was a landmark for medical education settings. The guidelines were reviewed years later in 2014 to emphasize other important domains for recently graduated general physicians, such as mental health, urgency, and emergencies.[17,18]

However, the equitable division of the number of items in general medical education assessments has not yet been adequately reappraised. The present results demonstrate that following this distribution may lead to suboptimal results in terms of assessment reliability, especially for internal medicine. This is why the consortium of public medical schools in the state of São Paulo changed the number of items across the content areas in progress testing in 2023. Consequently, the reliability index for internal medicine, which ranked highest, immediately increased. The increase in the number of items alone may have led to this result as the α coefficient is considerably influenced by the number of items.[19] While this may be true, it is only a part of the effect.

Following the well-accepted aphorism of biomedical research, increasing the sample size leads to more reliable results.[20] Twenty items do not seem sufficient to adequately sample all the content that medical students should know about internal medicine. In addition, if internal consistency refers to the extent to which different items measure the same construct, it is worth imagining how distant an item about acute myocardial infarction may be from another item about osteoarthritis. In contrast, for obstetrics, an item addressing gestational hypertension is expected to be similar to an item referring to HELLP syndrome (i.e., these two items probably measure the same construct).
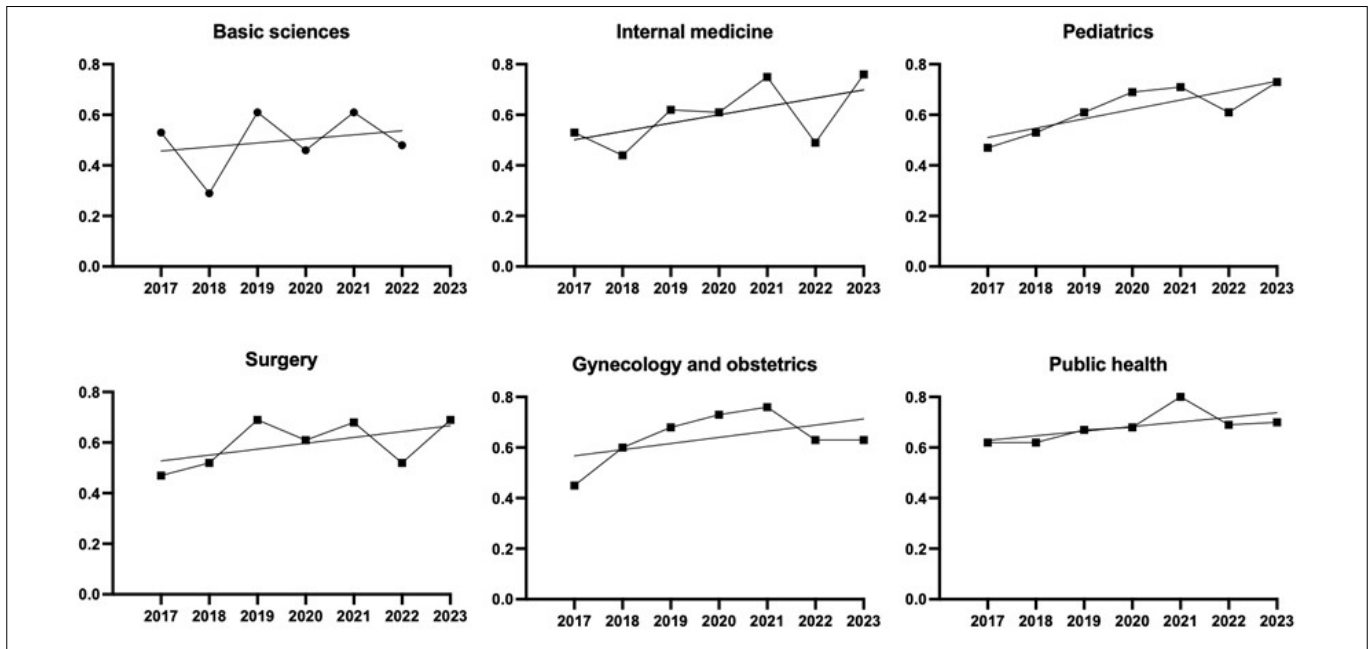
Figure 2. Alpha coefficient values (y axis) for each content area of progress testing between 2017 and 2023, with their respective trend lines. Basic sciences had the lowest intercept value and only pediatrics had a statistically significant upward trend.

Table 2. Linear regression models of temporal trends for each content area's reliability index

Area	Intercept	Slope	R^2	P value
Basic sciences	0.441	1.6%	0.063	0.632
Internal medicine	0.469	3.3%	0.332	0.176
Pediatrics	0.473	3.7%	0.691	**0.020**
Surgery	0.504	2.3%	0.288	0.214
Gynecology and obstetrics	0.543	2.4%	0.266	0.236
Public health	0.610	1.8%	0.420	0.116

The intercept refers to the point at which the regression line crosses the y axis. The slope refers to the annual percentage change. R^2 is the determination coefficient: the closer to "1," the better the adjustment of data to the regression. All areas showed a positive trend towards improvement in the reliability index, with a significant trend for Pediatrics (P < 0.05).

Incorporating basic sciences into the other content areas was a necessary adjustment to the test. As initially designed, progress tests must be structured at the graduate level, and therefore, writing items for "*first year students to get right*" is not adequate.[21] Therefore, items testing knowledge about basic sciences should be framed at the functional level of a recently graduated physician. Increasing the number of items in all content areas would be ideal; however, the extent to which tiredness undermines students' performance remains unclear. Therefore, we retained all 120 items.

We observed the best reliability indices for all six content areas in 2019, 2021, and 2023, and this is no coincidence. In those years, the progress test included pre-tested items, that is, items that had been previously used and selected based on their good psychometric properties. This finding highlights the benefits that medical schools can derive by supporting their faculty's ability to write good items that obtain better assessments[22]; this is especially important for basic sciences, whose faculties are lesser experienced

with writing clinical-based items.[23] Good items have higher taxonomic levels and better psychometric behavior, and therefore, they provide more reliable results.[24,25] Thus, better assessment is the key to improving medical education as it steers learning along the right path.[26] Moreover, we observed that the α coefficients of all content areas has been increasing over the years, which indicates that this consortium of public medical schools in the state of São Paulo has improved the quality of the progress test examination.

Previous studies have highlighted the need to better explore the validity of assessments in medical education.[27,28] In this context, validity measures the truth of the inferences made from the results of an examination. Reliability is not sufficient to guarantee validity but is a condition for it.[29] Our study focused on a particular domain of validity—that is, a psychometric focus on reliability. Nevertheless, our results call attention to the importance of a broader perspective on medical education assessment and the need to reappraise content sampling of medical knowledge.

REFERENCES

1. Ministério da Educação. Resolução n° 01, de 14 de agosto de 2000. Available from: https://www.gov.br/mec/pt-br/residencia-medica/ementario-da-legislacao-da-residencia-medica. Accessed in 2023 (Nov. 01)

2. Bicudo AM, Hamamoto Filho PT, Abbade JF, Hafner MLMB, Maffei CML. Consortia of cross-institutional Progress Testing for all medical schools in Brazil. Rev Bras Educ Med. 2019;43(4):151-6. https://dx.doi.org/10.1590/1981-52712015v43n4RB20190018.

3. Zanolli MB, Streit DS, Maciel DT, et al. Differences in clerkship development between public and private Brazilian medical schools: an overview. BMC Med Educ. 2020 Sep 21;20(1):316. PMID: 32957972; https://doi.org/10.1186/s12909-020-02193-3.

4. Tio RA, Schutte B, Meiboom AA, Get al. The progress test of medicine: the Dutch experience. Perspect Med Educ. 2016;5(1):51-5. PMID: 26754310; https://doi.org/10.1007/s40037-015-0237-1.

5. Nouns ZM, Georg W. Progress testing in German speaking countries. Med Teach. 2010;32(6):467-70. PMID: 20515374. https://doi.org/10.3109/0142159x.2010.485656.

6. United States Medical Licensing Examination. Step 2 CK Content Outline & Specifications. Available from: https://usmle.org/prepare-your-exam/step-2-ck-materials/step-2-ck-content-outline-specifications. Accessed in 2023 (Nov. 01)

7. Downing SM. Reliability: on the reproducibility of assessment data. Med Educ. 2004;38(9):1006-12. PMID: 15327684. https://doi.org/10.1111/j.1365-2929.2004.01932.x.

8. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16(3):297-334. https://dx.doi.org/10.1007/BF02310555.

9. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ. 2011;2:53-55. PMID: 28029643; https://doi.org/10.5116/ijme.4dfb.8dfd.

10. Cecilio-Fernandes D, Bicudo AM, Hamamoto Filho PT. Progress testing as a pattern of excellence for the assessment of medical students' knowledge - concepts, history, and perspective. Medicina (Ribeirão Preto). 2021;54(1):e-173770. http://dx.doi.org/10.11606/issn.2176-7262.rmrp.2021.173770.

11. Troncon LEA, Elias LLK, Osako MK, et al. Reflections on the use of the Progress Test in the programmatic student assessment. Rev Bras Educ Med. 2023;47(2):e076. https://dx.doi.org/10.1590/1981-5271v47.2-2022-0334.ing.

12. Tomic ER, Martins MA, Lotufo PA, Benseñor IM. Progress testing: evaluation of four years of application in the school of medicine, University of São Paulo. Clinics (Sao Paulo). 2005;60(5):389-96. PMID: 16254675. https://doi.org/10.1590/s1807-59322005000500007.

13. Sakai MH, Ferreira Filho OF, Almeida MJ, Mashima DA, Marchese MC. Progress testing and course evaluation: ten years of experience in the undergraduate medical course at the State University of Londrina. Rev Bras Educ Med. 2008;32(2):254-63. https://dx.doi.org/10.1590/S0100-55022008000200014.

14. Machado JLM. A participação da ANDES-SN na Cinaem: o processo de transformação das escolas médicas brasileiras. Interface (Botucatu). 1999;3(5):177–80. https://doi.org/10.1590/S1414-32831999000200025.

15. Falavigna A, Canabarro CT, Medeiros GS. Health system and medical education in Brazil: history, principles, and organization. World Neurosurg. 2013;80(6):723-7. PMID: 23333994. https://doi.org/10.1016/j.wneu.2013.01.079.

16. Ministério da Educação. Resolução CNE/CES n° 4, de 7 de Novembro de 2001. Brasilia: Ministério da Educação; 2001. Available from: http://portal.mec.gov.br/cne/arquivos/pdf/CES04.pdf. Accessed in 2023 (Nov. 01)

17. Ministério da Educação. Resolução CNE/CES n° 3, de 20 de Junho de 2014. Brasilia: Ministério da Educação; 2014. Available from: http://portal.mec.gov.br/cne/arquivos/pdf/CES04.pdf. Accessed in 2023 (Nov. 01)

18. Antunes Dos Santos R, Nunes MDPT. Medical education in Brazil. Med Teach. 2019;41(10):1106-1111. PMID: 31282823. https://doi.org/10.1080/0142159x.2019.1636955.

19. Agbo AA. Cronbach's alpha: review of limitations and associated recommendations. J Psychol Africa. 2010; 20(2):233-39. https://dx.doi.org/10.1080/14330237.2010.10820371.

20. Charter RA. Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. J Clin Exp Neuropsychol. 1999;21(4):559-66. PMID: 10550813. https://doi.org/10.1076/jcen.21.4.559.889.

21. Schuwirth LW, van der Vleuten CP. The use of progress testing. Perspect Med Educ. 2012;1(1):24-30. PMID: 23316456; https://doi.org/10.1007/s40037-012-0007-2.

22. Hamamoto Filho PT, Bicudo AM. Improvement of faculty's skills on the creation of items for progress testing through feedback to item writers: a successful experience. Rev Bras Educ Med. 2020;44(1):e018. https://dx.doi.org/10.1590/1981-5271v44.1-20190130.ING.

23. Feliciano CS, Elias LLK, Osako MK, Guimarães FS, Troncon LEA, Bollela VR. Workshop on item writing to assess basic sciences in health professional education: an experience report. Rev Bras Educ Med. 2023;47(2):e067. https://doi.org/10.1590/1981-5271v47.2-2022-0297.

24. Cecilio-Fernandes D, Kerdijk W, Bremers AJ, Aalders W, Tio RA. Comparison of level of cognitive process between case-based items and non-case-based items of the interuniversity progress test of medicine in the Netherlands. J Educ Eval Health Prof. 2018;15:28. PMID: 30541188; https://doi.org/10.3352/jeehp.2018.15.28.

25. Hamamoto Filho PT, Silva E, Ribeiro ZMT, Hafner MLMB, Cecilio-Fernandes D, Bicudo AM. Relationships between Bloom's taxonomy, judges' estimation of item difficulty and psychometric properties of items from a progress test: a prospective observational study. Sao Paulo Med J. 2020;138(1):33-39. PMID: 32321103; https://doi.org/10.1590/1516-3180.2019.0459.r1.19112019.

26. Norman G, Neville A, Blake JM, Mueller B. Assessment steers learning down the right road: impact of progress testing on licensing examination performance. Med Teach. 2010;32(6):496-9. PMID: 20515380. https://doi.org/10.3109/0142159x.2010.486063.

27. Kane MT. Validating the interpretations and uses of test scores. J Educ Meas. 2013;50(1):1-73. https://dx.doi.org/10.1111/jedm.12000.

28. Boursicot K, Kemp S, Wilkinson T, et al. Performance assessment: Consensus statement and recommendations from the 2020 Ottawa Conference. Med Teach. 2021;43(1):58-67. PMID: 33054524. https://doi.org/10.1080/0142159x.2020.1830052.

29. McLaughlin K, Lemaire J, Coderre S. Creating a reliable and valid blueprint for the internal medicine clerkship evaluation. Med Teach. 2005;27(6):544-7. PMID: 16199363. https://doi.org/10.1080/01421590500136113.

Address for correspondence:
Pedro Tadao Hamamoto Filho
Departamento de Neurologia, Psicologia e Psiquiatria, Faculdade de Medicina de Botucatu,
Universidade Estadual Paulista (UNESP)
Distrito de Rubião Jr, s/n - (SP), Brasil.
CEP 18618-686
Tel: +55 14 3880-1220
E-mail: pedro.hamamoto@unesp.br