https://doi.org/10.18222/eae.v34.9220

LONGITUDINAL ASSESSMENT OF MEDICAL STUDENTS: IS PROGRESS TEST APPROPRIATE?

CARLOS EDUARDO ANDRADE PINHEIRO[I]
DIOGO ONOFRE DE SOUZA[II]

[I] Universidade Federal de Santa Catarina (UFSC), Florianópolis-SC, Brazil; *ceapinheiro1@gmail.com*
[II] Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre-RS, Brazil; *diogo.bioq@gmail.com*

ABSTRACT

The purpose of this article is to assess whether the Progress Test is appropriate to evaluate programs and students during different stages of medical studies. The characteristics of the items and reliability of three previously applied progress tests were analyzed. For second-year students, 76.4% of the questions demonstrated poor quality (biserial < 0.2). This percentage decreased to 47.7% in the fourth year and to 25.3% in the sixth year. The test's reliability, measured by Cronbach's alpha, was only 0.60 for second year students and increased to 0.76 in the fourth year and to 0.87 for sixth-year students. The current form of the Progress Test showed low and unacceptable reliability for second-year students, reasonable for the fourth year, and excellent for the sixth year. An improvement of this longitudinal assessment is proposed.

KEYWORDS EDUCATION EVALUATION • MEDICAL EDUCATIONAL MEASUREMENT • EXTERNAL EVALUATION.

HOW TO CITE:

Pinheiro, C. E. A., & Souza, D. O. de. (2023). Longitudinal assessment of medical students: Is Progress Test appropriate? *Estudos em Avaliação Educacional*, *34*, Article e09220. https://doi.org/10.18222/eae.v34.9220

AVALIAÇÃO LONGITUDINAL DE ESTUDANTES DE MEDICINA: O TESTE DE PROGRESSO É APROPRIADO?

RESUMO

O artigo objetiva aferir se o Teste de Progresso é apropriado para avaliar cursos e estudantes em diferentes fases do curso de medicina. Analisam-se as características das questões e a confiabilidade de três testes de progresso já aplicados. Constatou-se que, para os estudantes do 2º ano, 76,4% das questões se mostraram de qualidade pobre (bisserial < 0,2); diminuindo para 47,7% no 4º ano e para 25,3% no 6º ano. A confiabilidade dos testes, pelo alfa de Cronbach, foi de somente 0,60 para os alunos do 2º ano, aumentando para 0,76 para os do 4º ano e 0,87 para os alunos do 6º ano. A forma atual do Teste de Progresso mostrou confiabilidade baixa e inaceitável para os alunos do 2º ano, razoável para os do 4º e ótima para os alunos do 6º ano. Um aperfeiçoamento dessa forma de avaliação longitudinal é proposto.

PALAVRAS-CHAVE AVALIAÇÃO DA EDUCAÇÃO • EDUCAÇÃO MÉDICA • AVALIAÇÃO EXTERNA.

EVALUACIÓN LONGITUDINAL DE ESTUDIANTES DE MEDICINA: ¿LA PRUEBA DE PROGRESO ES APROPIADA?

RESUMEN

El artículo tiene el propósito de verificar si la Prueba de Progreso es apropiada para evaluar cursos y estudiantes en distintas fases del curso de medicina. Se analizan las características de las preguntas y la confiabilidad de tres pruebas de progreso ya aplicadas. Se constató que, para los estudiantes de 2º año, el 76,4% de las preguntas se mostraron de baja calidad (biserial < 0,2), reduciéndose a 47,7% en 4º año y a un 25,3% en 6º año. La confiabilidad de las pruebas, por alfa de Cronbach, fue de tan solo 0,60 para los alumnos de 2º año, y aumentó a 0,76 para los de 4º año y a 0,87 para los estudiantes de 6º año. La forma actual de la Prueba de Progreso mostró confiabilidad baja e inaceptable para los alumnos de 2º año, razonable para los de 4º y excelente para los de 6º. Se propone que se perfeccione dicha forma de evaluación longitudinal.

PALABRAS CLAVE EVALUACIÓN DE LA EDUCACIÓN • EDUCACIÓN MÉDICA • EVALUACIÓN EXTERNA.

Received on: DECEMBER 16, 2021
Approved for publication on: MARCH 22, 2023

INTRODUCTION

Although scholars concur that longitudinal knowledge assessment is advantageous for medical education (Albanese & Case, 2016; Cecilio-Fernandes et al., 2021; Vleuten et al., 2018; Wrigley et al., 2012), one of the great challenges facing current medical education is preparing assessments that can simultaneously i) measure the knowledge acquired by students throughout their studies (summative assessment), ii) be an instrument for the students to adjust their studies (formative assessment), and iii) help the professor and the educational institution to understand (informative assessment) and to improve knowledge acquisition (Pugh & Regehr, 2016). One type of extensive assessment of cognitive knowledge that can perform these three functions, the Progress Test (PT), has been used in medical schools. It is periodically conducted and applied on the same day to all medical undergraduates, targeting the knowledge level expected for final-year students. However, there is uncertainty whether the current structure of PT is suitable to assess students in the early and intermediate stages of the program (Albanese & Case, 2016; Henning et al., 2017).

PT arose in the wake of the constructivist pedagogy theory, which proposes as one of its principles that students are the builders of their own knowledge; the professor is no longer the granter of knowledge but a partner in the search for knowledge. This great change resulted in the use of new forms of approaching teaching and learning in medical education, called active methodologies, such as Problem-Based Learning (PBL), Team-Based Learning (TBL), etc. These new ways of teaching have generated the need to create new forms of assessing students and schools (Cecilio-Fernandes et al., 2021; Heeneman et al., 2017; Reberti et al., 2020; Vleuten & Schuwirth, 2019).

Schools with traditional pedagogy, which do not use active methodologies, tend to assess student learning in each one of the disciplines or areas of knowledge separately. At the year-end exams, it is common for students to take many tests during the same week. They frequently study for each one of the tests at the last moment, adopting lower-order cognitive strategies such as cramming or memorizing by rote (Pugh & Regehr, 2016). This kind of assessment promotes uncontextualized and fragmented learning that is forgotten after a short time. The PT, in contrast, because it covers a much broader content, is integrative and provides timely feedback, discourages the aforementioned strategy and stimulates a deeper, contextualized, and lasting knowledge (Vleuten et al., 2004). Assessments with wide-ranging content that are systematically and regularly applied throughout the program allow students to verify their strengths and weaknesses of knowledge and potentially provide motivation and direction for future learning (Epstein, 2007; Heeneman et al., 2017; Tavakol & Dennick, 2017).

Used in several countries since the 1970s, the PT is routinely applied in European and North American medical schools 2–4 times a year, with varying numbers of questions (Blake et al., 1996; Vleuten et al., 2004). It emerged as an assessment more aligned with the problem-based curriculum and as a tool to compare the knowledge acquired using this type of curriculum with the traditional ones (Cecilio-Fernandes et al., 2021). Currently, in the Netherlands, the test consists of 200 multiple-choice questions, with two-five options and penalties for wrong answers.

The items are built and revised by professors from different universities to avoid representing a vision and experience from an individual professor. In most cases, the questions are clinical vignettes, not long ones, that require a type of knowledge and logical reasoning involving higher-order cognitive processes. They are a matrix that comprises all knowledge areas expected by the end of the course. On a set date and time, all medical students take the same exam. Afterward, they receive their test scores individually, their acquired knowledge compared to previous year(s), and different ways of feedback.

In Brazil, since 2001, several consortiums of medical schools have been created to elaborate assessments with PTs. However, only one test of 120 simple multiple-choice questions (MCQs), with four answer options without penalization for wrong answers, is administered each year. With the support of the Brazilian Medical Education Association (ABEM), there are currently thirteen regional consortia that administer the PT in their medical schools. In 2015, 40% of the Brazilian medical schools had joined a PT consortium (Bicudo et al., 2019; Rosa et al., 2017; Sakai et al., 2008; Sartor et al., 2020). The organization of schools in the form of consortiums for the PT promotes integration among schools, qualifies professors to elaborate items (questions), and reduces costs associated with the process of test elaboration, printing, and correction (Hamamoto & Bicudo, 2020b; Vleuten et al., 2018).

Brazil has 357 medical schools, with 37,823 slots for new students in 2020; in 2001, 11,541 new slots were offered (Scheffer et al., 2020). This sharp increase ignited the discussion about creating license exams for practicing medicine, in view of the need to create new external assessment formats to ensure the quality of schools and graduates (Bica & Kornis, 2020; Troncon, 2019). The PT in Brazil is a form of longitudinal and essentially formative external assessment, offering different feedback opportunities. Regardless of the type of curriculum adopted by the school (traditional or PBL), the PT scores hardly ever affect regular student grades or whether the student will pass their year-end tests of final exams.

In some European and American medical schools, the PT results also add scores that will determine whether the student can progress through the program (summative assessment), or they are classified according to their performance on

the exam (Albanese & Case, 2016; Blake et al., 1996; Heeneman et al., 2017). These PTs must therefore be reliable (Downing, 2004; Kibble, 2017). In Holland, four national PTs are administered each year to all medical students in order to increase assessment reliability (Wrigley et al., 2012); this longitudinal knowledge assessment throughout the program also replaces the license exam at the end of medical programs (Vleuten et al., 2004). However, these longitudinal, nationwide tests at such frequencies involve costs that many countries, such as Brazil, can hardly bear.

Both in Brazil and abroad, the schools receive their students' general scores and the scores for the different knowledge areas covered in the medical curriculum, such as basic sciences, collective health, pediatrics, primary care, etc. By comparing students' scores, both globally and for the different knowledge areas, with the mean for the group of PT-participant schools, it is possible for the schools to identify which areas correspond to their strengths and weaknesses. Based on this information, the schools can alter their curriculum, processes and teaching and learning strategies so that in the following years they can evaluate the impact of these changes (Rosa et al., 2017).

PT organizers, in all parts of the world where it is applied, after the exam is employed, analyze the test's validity and reliability, its pedagogical aspects, and the psychometrics of each item. They select some questions and modify others to build a question database. The emergence of computerized adaptive testing made it possible for questions to become more or less difficult as the student answers the test, depending on their skills and knowledge. The existence of an item bank, with questions previously tested and calibrated, is fundamental for this form of assessment. An international computerized PT is already available and can provide results with a high level of reliability, using a smaller number of questions (Cecilio-Fernandes, 2019; Collares & Cecilio-Fernandes, 2019).

In Brazil, the Ministry of Education has been applying a nationwide exam, the "Exame Nacional de Desempenho dos Estudantes (ENADE)", to all final-year undergraduates, including medical students, since 1996. This exam was changed in 2004 to include first-year students in order to determine programs' contribution to student learning. However, after ENEM – the exam applied to senior high school students and used for admission to public universities and as a quality reference for a number of public higher education inclusion policies – became national in scope, this initiative was abandoned, and now only graduating students are required to take it. Since this forty-question exam (with only thirty questions directly addressing the professional field) is applied to each career only once every three years, it has little formative impact and is clearly inappropriate for assessing students' learning progress (Ristoff, 2022). This has led medical schools to move forward with other forms of assessment such as the PT.

To improve the reliability of PTs and to overcome the economic issues related to running multiple tests each year, the development of a better-quality PT is a promising approach to improve the assessment of medical students' knowledge (Aubin et al., 2020; Sahoo & Singh, 2017). Therefore, this study aimed to analyze the reliability of PTs administered in Brazil for students at different stages of their studies (2nd, 4th and 6th years). Based on our results, we propose a new format of longitudinal knowledge assessment, called the Customized Progress Test (CPT), which can evaluate medical schools and the progress of students throughout their studies, serving as an alternative to license exams applied only at the end of programs.

OBJECTIVES

1. To evaluate whether current Brazilian Progress Tests are appropriate to assess schools and students' knowledge in the initial, intermediate, and final stages of medical programs.
2. To build relevant knowledge so as to inform a proposal for improving the PT, thereby collaborating to create a form of longitudinal assessment that enhances the assessment of medical students' knowledge in all program years across Brazil

METHODS

Design, Location and Participants

The results were obtained from three previously applied PTs which covered the expected knowledge for 6th-year students and were administered to students in three different years of the medical program. The characteristics of the items (difficulty and discrimination index) and the reliability of the tests were calculated.

The analyzed PTs were prepared by a consortium of ten medical schools with support from the Brazilian Medical Schools Association (ABEM) – South Branch II. The PTs were administered in 2015, 2016, and 2017 and taken simultaneously by all students of the ten schools. However, only the results for students in the 2nd, 4th and 6th years of medical programs were analyzed, instead of all years, because of a not yet implemented proposal to apply a nationwide longitudinal knowledge assessment for students only in those school years (Ministério da Educação, 2016).

In the analyzed PTs, the number of 2nd-year students was the highest and that of 6th-year ones the lowest, since three of the participant schools were new, and the other two had increased their number of students just before the tests

(Scheffer & Dal Poz, 2015). In Brazil, medical programs in all higher education institutions last for 6 years.

Characterization of the Applied Progress Tests (PTs)

The three PTs were designed to evaluate all students based on the expected knowledge for final-year (6th-year) students, with the content of the tests based on a matrix of themes from the Ministry of Education's "National Guidelines for Medical Programs" (Resolução n. 3, 2014). Each PT had 120 simple MCQs with four answer options, of which only one was correct, and without any negative scoring for wrong answers. All three PTs had twenty items from each one of six major areas: basic sciences, clinical medicine, surgery, gynecology/obstetrics, pediatrics, and public health.

Estimated Variables and Statistical Analysis

In the three PTs, the following variables were analyzed:

i) item difficulty;
ii) item discrimination index;
iii) test reliability.

Item difficulty was determined by the percentage of correct answers for each question: those with less than 45% correct answers were considered difficult, those having between 45% and 80% were considered average, and those with more than 80% correct answers were considered easy.

The item discrimination index was determined by a point-biserial correlation coefficient, which measures the correlation between the correctness or error in each item (question) to the final test score and indicates the quality of each question. Questions with a point-biserial correlation coefficient below 0.2 are considered to have poor discrimination, indicating that they must be reviewed or eliminated, while those with a coefficient above 0.2 are considered acceptable or good (De Champlain, 2010; Walsh et al., 2018).

Test reliability was measured using Cronbach's alpha coefficient, which measures the internal consistency of the test; reliability is the probability that the test score represents the actual knowledge of the person being evaluated – it should be at least 0.9 for a high-stake assessment; values between 0.89 and 0.8 are good, and 0.7 is considered the minimum acceptable for broad-spectrum educational assessments (Downing, 2004; Kibble, 2017; Vleuten & Schuwirth, 2005).

Statistical Analysis

Item difficulty and the item discrimination index for each question were measured in phase 1 of the BILOG-MG software (version 3.0.2327.2 – Scientific

Software International, Inc.) for Windows. The comparison between the distribution averages of question difficulty and discrimination was carried out by one-way analysis of variance (ANOVA), and the reliability of PTs by Cronbach's alpha and two-way ANOVA without replication; these comparisons and the reliability ranges (α <0.05) were calculated with Microsoft Excel Professional 2016.

The scores presented are the percentage of correct answers for each student in the test.

The consortium organizing the PTs allowed this study under the condition that the identities of schools and students were kept confidential.

RESULTS

The number of students participating in the PTs (2nd, 4th and 6th-year students) and the average scores they obtained in the three analyzed years (2015, 2016, and 2017), are shown in Table 1.

TABLE 1

Number of students (N) in different program years, and average scores obtained on the three Progress Tests carried out in ten medical schools in Brazil's South Region – 2015, 2016 e 2017

YEAR OF EXAMS	STUDENTS					
	2nd YEAR		4th YEAR		6th YEAR	
	N	SCORES	N	SCORES	N	SCORES
2015	716	42.5	517	52.0	386	60.4
2016	822	39.8	539	49.5	411	58.3
2017	821	43.0	635	53.8	502	62.7
MEANS*		41.8		51.8		60.5
CI 95%		(40.2-43.4)		(49.8-53.7)		(58.4-62.5)

Source: The authors.

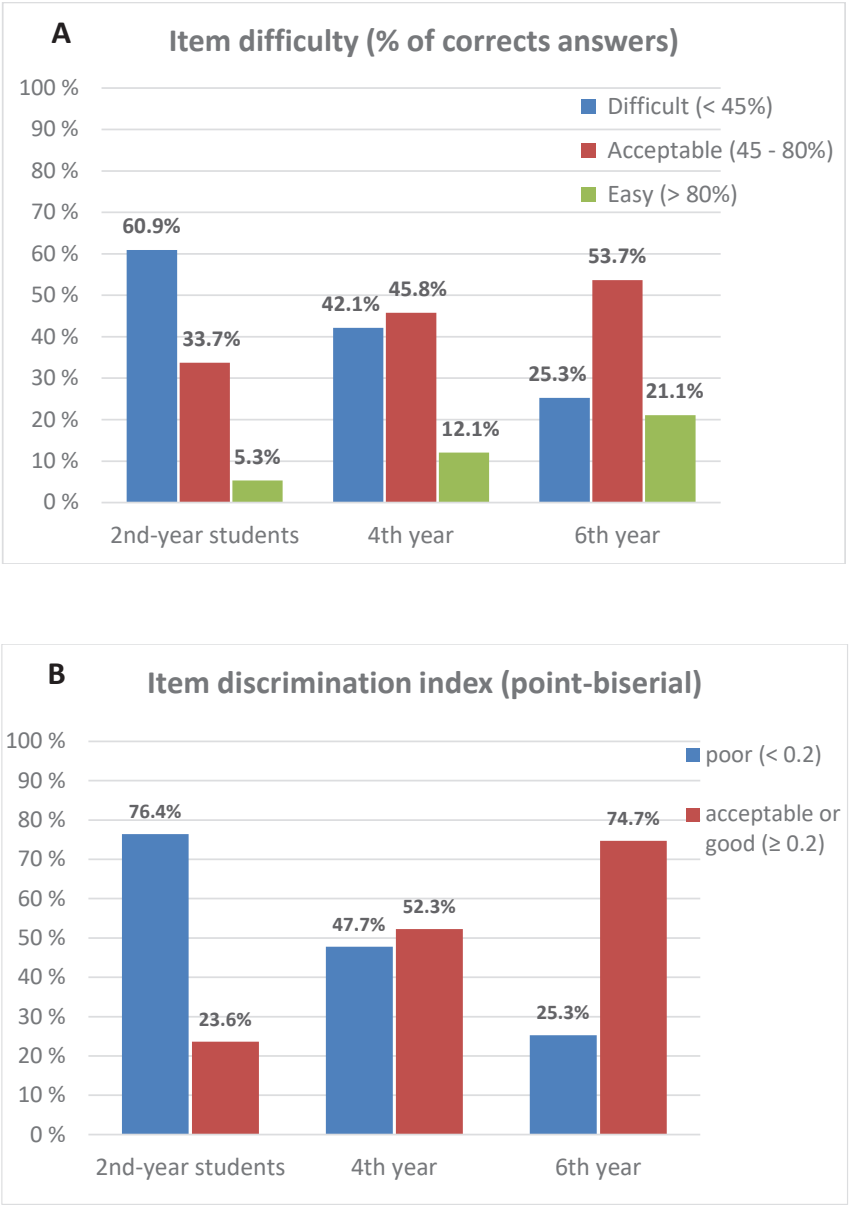
* ANOVA: $p < 0.0001$.

The students' mean scores progressively increased from the 2nd year (41.8) to the 4th year (51.8) and to the 6th year (60.5); the differences among averages were statistically significant.

The items' difficulty and discrimination mean values are shown in Figure 1.

FIGURE 1

Mean difficulty and discrimination values for PT items measured for students in different program years – means for 2015, 2016 and 2017 – ten medical schools in Brazil's South Region



Source: The authors.

The percentage of items considered difficult decreased from 60.9% (for 2nd-year students) to 42.1% and 25.3% (for 4th and 6th-year students, respectively), whereas that of questions considered easy increased from 5.3% (for 2nd-year students) to 12.1% and 21.1% (for 4th and 6th-year students, respectively) (Figure 1A).

The percentage of items with poor discrimination (bad quality) decreased from 2nd-year students (76.4%) to 4th and 6th-year ones (47.7% and 25.3%, respectively). All these comparisons between question characteristics for students in different program years proved to be significant ($p < 0.001$) by ANOVA (Figure 1B).

The reliability of the three PTs for the three program years and the years they were administered are shown in Table 2.

TABLE 2

Reliability of the three Progress Tests, measured by Cronbach's alpha, for the different program years – ten medical schools in Brazil's South Region – 2015, 2016 and 2017

YEAR OF EXAMS	STUDENTS		
	2nd YEAR	4th YEAR	6th YEAR
2015	0.64	0.75	0.86
2016	0.52	0.76	0.87
2017	0.65	0.77	0.87
MEANS*	0.60	0.76	0.87
CI 95%	(0.54-0.67)	(0.75-0.77)	(0.86-0.87)

Source: The authors.

* ANOVA: $p < 0.0001$.

The mean reliability of the three PTs, measured by Cronbach's alpha, increased from 2nd-year students (0.60) to 4th and 6th-year ones (0.76 and 0.87, respectively). The differences between the mean values were significant ($p < 0.001$), and there was no overlap of confidence intervals for the mean values in any of the comparisons of PTs' reliabilities between students in different program years.

DISCUSSION

The mean score differences between 2nd, 4th, and 6th-year students, in all PTs described in Table 1, show a similar increase in the three years analyzed and are in accordance with the national (Bicudo et al., 2019) and international (Wrigley et al., 2012) literature.

In order to contribute to understanding the current PTs' assessment approach, this study evaluated i) item difficulty, ii) item discrimination index, and iii) test reliability of PTs administered to 2nd, 4th and 6th-years students from 10 Brazilian medical schools in 2015, 2016, and 2017. The percentage of items considered difficult decreased, and those considered easy increased from the 2nd to the 4th and 6th years (Figure 1A). The discrimination index value for questions considered poor decreased, and for those considered acceptable or good, it increased from the 2nd to the 4th and 6th years (Figure 1B). The discrimination index determined through point-biserial correlation is a good indicator of an item's quality (Tavakol & Dennick, 2017). Point-biserial values below 0.2 suggest that the question is not able to discriminate between students with different levels of knowledge, and thus must

be considered cautiously (De Champlain, 2010; Kibble, 2017; Pasquali, 2004; Tavakol & Dennick, 2013). The 76.4% percentage of poor-quality items (point-biserial < 0.2) for 2nd-year students, found in this study, probably occurs because the PT is usually designed to target the knowledge level expected for final-year students, influencing the test´s reliability.

The reliability of PTs is influenced by the quality and number of questions and how frequently the test is administered (Kibble, 2017; Wrigley et al., 2012). The mean reliability of the PTs in this study (Table 2), measured by Cronbach's alpha, was 0.60 for 2nd-year students, which is considered low and unacceptable for moderate or even low-stakes assessments (Downing, 2004; Kibble, 2017). The mean reliability increased to 0.76 for 4th-year students, a number considered reasonable, and to 0.87 for 6th-year students, a number considered close to ideal. This high reliability for the 6th year suggests that the PT exams are reaching the expected consistency for this kind of assessment for the students in the final year of the medical program (Downing, 2004; Kibble, 2017; Pasquali, 2004). In a previously described PT study (Wrigley et al., 2012), similar values for 2nd and 4th-year students were observed. In the present study, the PTs' reliability of 0.60 for 2nd-year students indicates that the PT scores presented a larger random error for this program year (Tavakol & Dennick, 2011). Such a high number of errors measured in a single annual PT with 120 MCQs, as in this study, makes the test´s usefulness impractical for the current assessment of 2nd-year medical students in Brazil (Albanese & Case, 2016; Downing, 2004; Kibble, 2017). Thus, this low reliability confirms the difficulty of using this format of PT in Brazil as a longitudinal assessment aimed as an alternative to a national license exam for professional medical practice.

The standard performance values used in this study for item difficulty, item discrimination index, and test reliability can be questioned. Some authors consider items with less than 20% (Aubin et al., 2020) or 30% (De Champlain, 2010; Kheyami et al., 2018; Primi, 2012; Sahoo & Singh, 2017) correct answers as difficult. These values are normally used when there is a penalty for incorrect answers (formula scoring), in order to avoid guessing. Regarding the discrimination index, some authors consider items with a point-biserial correlation coefficient of 0.25 or above (Tavakol & Dennick, 2017) as good. However, it is almost unanimous that a test with a reliability value below 0.70, as observed for 2nd-year students in this study, cannot be used to assess the academic development of students.

In some European locations, where PT emerged over fifty years ago, the purpose was to evaluate curriculum changes and create examinations aligned with PBL curriculums. Interestingly, it was initially used only as a formative assessment. Gradually, it substituted the traditional forms of summative assessment. With the increase in the frequency of annual examinations in order to increase the reliability

of results, in some places, the PT became the only tool to evaluate knowledge or its acquisition. The PT has been consistently one of the components of the assessment system, which also analyzes skills and competencies. However, this is not currently used in Brazil, where the PT started to be employed less than twenty years ago and is mainly used as a formative tool.

The discussion about the need to use one PT each semester (twice a year) and about ways of integrating the results into summative assessments has begun in our consortiums. However, both in schools with modern curriculums, pedagogically aligned with the PT, and in schools with traditional curriculums, this discussion needs to be deepened. Moreover, it seems necessary to increase student adherence and commitment, especially since the initial phases, and to avoid guessing behavior. However, this issue has not been studied or described in the literature regarding PT and medical education in Brazil.

In general, the PTs are developed and employed by a group of schools. This collective initiative reduces the costs of PT employment and analysis and increases the quality of questions. In addition to the advantage for students, since they can rectify their journey early on, the PT also provides a comparative parameter to the institutions, which can analyze their strengths and weaknesses. In Brazil, the concern with results and their use as a tool for changes in schools seems to elicit more interest from private schools than public ones, as apparently occurs already with ENADE (Damas & Miranda, 2019).

In order to avoid the risks of ENADE, which creates a ranking and a rivalry for positions between schools and might serve as a parameter that can cover up issues, with PT, each school receives only its own results, and the average results for the group of participant schools serves as a comparison factor. In addition to fostering integration between the schools in the consortium, this environment of collaboration strengthens ABEM regional branches. Moreover, it stimulates the creation of an assessment research area and holds the potential to improve knowledge assessments and medical education (Hamamoto & Bicudo, 2020a).

The elaboration of tests with well-structured multiple-choice questions is considered adequate to assess the theoretical knowledge evolution of students, also called cognitive domain assessment (Aubin et al., 2020; Epstein, 2007; Vleuten & Schuwirth, 2019). This must partially explain why the PT is being increasingly applied in more countries (Vleuten et al., 2018) and in different undergraduate programs, such as dentistry (Ali et al., 2016; Oliveira et al., 2020), pharmacy (Albekairy et al., 2021) and veterinary (Herrmann et al., 2020), as well as in postgraduate medical training (Rutgers et al., 2018), both in Brazil and abroad (Alkhalaf et al., 2021; Sá et al., 2021).

Because of its wide content, the rise of PT discourages last-minute study and encourages continuous study, thus representing an advancement in educational

assessments in the healthcare field (Pugh & Regehr, 2016). However, the most modern concepts of assessments suggest that each school should create not only one or two types of evaluation, but assessment systems that cover all theoretical knowledge domains, skills, and attitudes (Norcini et al., 2018). The same concepts should be applied to large-scale assessments.

Other essential domains for professional practice, such as communication skills, clinical skills, and attitudes should also be evaluated (Epstein, 2007; Vleuten, 2016). However, they require much more complex, laborious, and expensive evaluation formats, such as Objective Structured Clinical Examination (OSCE), Mini Clinical Evaluation Exercise (Mini-Cex), 360 Degree Feedback, etc. The application of these methodologies in large-scale external educational assessments deserves further study and analysis to become viable for use in Brazil.

If one wide, large-scale longitudinal assessment system is inadequate for determining whether a professional has enough knowledge and is able to practice medicine, even more so would be a single theoretical exam, such as the professional license exam, applied at the end of medical programs. Such an exam would neither directly impact schools nor improve medical education. It would be applied to test students, whether or not they acquired theoretical knowledge during their undergraduate studies, without having allowed them the possibility of remediation while in the program. Hence the need for reliable longitudinal assessments for students at different stages of the program.

Thus, considering only the assessment of theoretical knowledge, this study supports the perception that the current PT format, i.e., 120 MCQs, with a test administered once a year in Brazil, targeting the knowledge expected for 6th-year students, is unsuitable for assessing the knowledge acquired by medical students since the initial medical school years. The PT can and should be improved for students in their initial and intermediate stages.

This study is the first of a series that aims to investigate and propose a reliable longitudinal assessment that measures knowledge acquisition by students in all their study years in the medical program. Thus, we intend to apply a new type of PT called the Customized Progress Test (CPT). In this approach, the items for 2nd-year students will be divided as follows: 25% will cover the expected knowledge level for their year, 25% for the 4th-year level, and 50% for the 6th-year level; for 4th-year students, 50% of the questions will cover the expected knowledge level for their year and 50% for 6th-year level; and for 6th-year students, all questions will cover the expected knowledge level for their year, as is currently done with excellent reliability (Wrigley et al., 2012). It is expected that this approach will help to provide an assessment (1) with good reliability for all stages of the program and (2) which can be used as an evaluation with moderate–high stakes for examinees.

In its current state, the yearly longitudinal PT in medical programs is unreliable for evaluating the knowledge that medical students acquired over their medical program years. We hope that this study will reinforce the relevance of proposing the Customized Progress Test as an innovative strategy for formative and summative assessments, potentially contributing to significantly improve the evaluation of medical education in Brazil. Such an assessment as the one proposed by the CPT has the potential to be a good alternative to the proposed national licensing examination at the end of healthcare programs.

REFERENCES

Albanese, M., & Case, S. M. (2016). Progress testing: Critical analysis and suggested practices. *Advances in Health Sciences Education*, *21*(1), 221-234. https://doi.org/10.1007/s10459-015-9587-z

Albekairy, A. M., Obaidat, A. A., Alsharidah, M. S., Alqasomi, A. A., Alsayari, A. S., Albarraq, A. A., Aljabri, A. M., Alrasheedy, A. A., Alsuwayt, B. H., Aldhubiab, B. E., Almaliki, F. A., Alrobaian, M. M., Aref, M. A., Altwaijry, N. A., Alotaibi, N. H., Alkahtani, S. A., Bahashwan, S. A., & Alahmadi, Y. A. (2021). Evaluation of the potential of national sharing of a unified progress test among colleges of pharmacy in the Kingdom of Saudi Arabia. *Advances in Medical Education and Practice*, *12*, 1465-1475. https://doi.org/10.2147/amep.s337266

Ali, K., Coombes, L., Kay, E., Tredwin, C., Jones, G., Ricketts, C., & Bennett, J. (2016). Progress testing in undergraduate dental education: The Peninsula experience and future opportunities. *European Journal of Dental Education*, *20*(3), 129-134. https://doi.org/10.1111/eje.12149

Alkhalaf, Z. S. A., Yakar, D., Groot, J. C. de, Dierckx, R. A. J. O., & Kwee, T. C. (2021). Medical knowledge and clinical productivity: Independently correlated metrics during radiology residency. *European Radiology*, *31*(7), 5344-5350. https://doi.org/10.1007/s00330-020-07646-3

Aubin, A.-S., Young, M., Eva, K., & St-Onge, C. (2020). Examinee cohort size and item analysis guidelines for health professions education programs: A Monte Carlo simulation study. *Academic Medicine: Journal of the Association of American Medical Colleges*, *95*(1), 151-156. https://doi.org/10.1097/acm.0000000000002888

Bica, R. B. da S., & Kornis, G. E. (2020). Exames de licenciamento em medicina: Uma boa ideia para a formação médica no Brasil? *Interface – Comunicação, Saúde, Educação*, *24*, Artigo e180546. https://doi.org/10.1590/Interface.180546

Bicudo, A. M., Hamamoto, P., Filho, Abbade, J., Hafner, M. de L., & Maffei, C. (2019). Teste de Progresso em Consórcios para todas as escolas médicas do Brasil. *Revista Brasileira de Educação Médica*, *43*(4), 151-156. https://doi.org/10.1590/1981-52712015v43n4RB20190018

Blake, J. M., Norman, G. R., Keane, D. R., Mueller, C. B., Cunnington, J., & Didyk, N. (1996). Introducing progress testing in McMaster University's problem-based medical curriculum: Psychometric properties and effect on learning. *Academic Medicine*, *71*(9), 1002-1007. https://doi.org/10.1097/00001888-199609000-00016

Cecilio-Fernandes, D. (2019). Implementando o teste adaptativo computadorizado. *Scientia Medica*, *29*(3), Artigo e34432. https://doi.org/10.15448/1980-6108.2019.3.34432

Cecilio-Fernandes, D., Bicudo, A. M., & Hamamoto, P. T., Filho. (2021). Progress testing as a pattern of excellence for the assessment of medical students' knowledge – Concepts, history, and perspective. *Medicina, 54*(1), Article e-173770. https://doi.org/10.11606/issn.2176-7262.rmrp.2021.173770

Collares, C. F., & Cecilio-Fernandes, D. (2019). When I say... computerised adaptive testing. *Medical Education, 53*(2), 115-116. https://doi.org/10.1111/medu.13648

Damas, B. R., & Miranda, G. J. (2019). Preparação da Instituição para o Enade: Importa? In *Anais do 3. Congresso UFU de Contabilidade*. UFU. https://eventos.ufu.br/sites/eventos.ufu.br/files/documentos/030_artigo_completo.pdf

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education, 44*(1), 109-117. https://doi.org/10.1111/j.1365-2923.2009.03425.x

Downing, S. M. (2004). Reliability: On the reproducibility of assessment data. *Medical Education, 38*(9), 1006-1012. https://doi.org/10.1111/j.1365-2929.2004.01932.x

Epstein, R. M. (2007). Assessment in medical education. *New England Journal of Medicine, 356*(4), 387-396. https://doi.org/10.1056/nejmra054784

Hamamoto, P. T., Filho, & Bicudo, A. M. (2020a). Improvement of faculty's skills on the creation of items for progress testing through feedback to item writers: a successful experience. *Revista Brasileira de Educação Médica, 44*(1), Article e018. https://doi.org/10.1590/1981-5271v44.1-20190130.ING

Hamamoto, P. T., Filho, & Bicudo, A. M. (2020b). Implementation of the Brazilian National Network for Practices and Research with Progress Testing – BRAZ-NPT. *Revista Brasileira de Educação Médica, 44*(3), Letter to the editor e074. https://doi.org/10.1590/1981-5271v44.3-20200089

Heeneman, S., Schut, S., Donkers, J., Vleuten, C. van der, & Muijtjens, A. (2017). Embedding of the progress test in an assessment program designed according to the principles of programmatic assessment. *Medical Teacher, 39*(1), 44-52. https://doi.org/10.1080/0142159x.2016.1230183

Henning, M., Pinnock, R., & Webster, C. (2017). Does Progress Testing violate the principles of constructive alignment? *Medical Science Educator, 27*(4), 825-829. https://doi.org/10.1007/s40670-017-0459-4

Herrmann, L., Beitz-Radzio, C., Bernigau, D., Birk, S., Ehlers, J. P., Pfeiffer-Morhenn, B., Preusche, I., Tipold, A., & Schaper, E. (2020). Status quo of progress testing in veterinary medical education and lessons learned. *Frontiers in Veterinary Science, 7*, Article 559. https://doi.org/10.3389%2Ffvets.2020.00559

Kheyami, D., Jaradat, A., Al-Shibani, T., & Ali, F. A. (2018). Item analysis of multiple choice questions at the department of paediatrics, Arabian Gulf University, Manama, Bahrain. *Sultan Qaboos University Medical Journal, 18*(1), e68-e74. https://doi.org/10.18295/squmj.2018.18.01.011

Kibble, J. (2017). Best practices in summative assessment. *Advances in Physiology Education, 41*(1), 110-119. https://doi.org/10.1152/advan.00116.2016

Ministério da Educação. (2016). *Avaliação Nacional Seriada dos Estudantes de Medicina – Documento Básico*. Ministério da Educação.

Norcini, J., Anderson, M. B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., Hays, R., Palacios Mackay, M., Roberts, T., & Swanson, D. (2018). 2018 Consensus framework for good assessment. *Medical Teacher, 40*(11), 1102-1109. https://doi.org/10.1080/0142159x.2018.1500016

Oliveira, F. A. M., Martins, M. T., Ferraz, A. M. L., Júnior, Ribeiro, C. G., Oliveira, R. G. de, & Porto, F. R. (2020). Percepção dos acadêmicos de odontologia em relação ao Teste de Progresso. *Revista da Abeno, 20*(2), 26-37. https://doi.org/10.30979/rev.abeno.v20i2.821

Pasquali, L. (2004). *Psicometria dos testes na psicologia e na educação* (5a ed.). Vozes.

Primi, R. (2012). Psicometria: Fundamentos matemáticos da teoria clássica dos testes. *Avaliação Psicológica, 11*(2), 297-307.

Pugh, D., & Regehr, G. (2016). Taking the sting out of assessment: Is there a role for progress testing? *Medical Education, 50*(7), 721-729. https://doi.org/10.1111/medu.12985

Reberti, A. G., Monfredini, N. H., Ferreira, O., Filho, Andrade, D. F. de, Pinheiro, C. E., & Silva, J. C. (2020). Teste de Progresso na escola médica: Uma revisão sistemática acerca da literatura. *Revista Brasileira de Educação Médica, 44*(1), Artigo e015. https://doi.org/10.1590/1981-5271v44.1-20190194

Resolução n. 3, de 20 de junho de 2014. (2014). Institui Diretrizes Curriculares Nacionais do Curso de Graduação em Medicina e dá outras providências. *Diário Oficial da União*, Brasília, DF.

Ristoff, D. (2022). *Mitos e meias-verdades: A educação superior sob ataque*. Insular.

Rosa, M. I. da, Isoppoi, C., Cattaneoi, H., Madeirai, K., Adami, F., & Ferreira, O. F., Filho. (2017). O Teste de Progresso como indicador para melhorias em Curso de Graduação em Medicina. *Revista Brasileira de Educação Médica, 41*(1), 58-68. https://doi.org/10.1590/1981-52712015v41n1RB20160022

Rutgers, D., Raamt, F. van, Lankeren, W. van, Ravesloot, C., Gijp, A. van der, Ten Cate, T. J., & Schaik, J. van. (2018). Fourteen years of progress testing in radiology residency training: Experiences from The Netherlands. *European Radiology, 28*(5), 2208-2215. https://doi.org/10.1007%2Fs00330-017-5138-8

Sá, M. F. S. de, Romão, G. S., Fernandes, C. E., & Silva, A. L. da, Filho. (2021). The Individual Progress Test of Gynecology and Obstetrics Residents (TPI-GO): The Brazilian Experience by Febrasgo. *Revista Brasileira de Ginecologia e Obstetrícia, 43*(6), 425-428. https://doi.org/10.1055/s-0041-1731803

Sahoo, D., & Singh, R. (2017). Item and distracter analysis of multiple choice questions (MCQs) from a preliminary examination of undergraduate medical students. *International Journal of Research in Medical Sciences, 5*(12), 5351-5355. http://dx.doi.org/10.18203/2320-6012.ijrms20175453

Sakai, M. H., Ferreira, O. F., Filho, Almeida, M., Mashima, D., & Marchese, M. (2008). Teste de Progresso e avaliação do curso: Dez anos de experiência da medicina da Universidade Estadual de Londrina. *Revista Brasileira de Educação Médica, 32*(2), 254-263. https://doi.org/10.1590/S0100-55022008000200014

Sartor, L. B., Rosa, L., Rosa, M. I. da, Madeirai, K., Uggioni, M. L., Ferreira, O., Filho. (2020). Undergraduate Medical Student's Perception about the Progress Testing. *Revista Brasileira de Educação Medica, 44*(2), Artigo e062. https://doi.org/10.1590/1981-5271v44.2-20190286.ING

Scheffer, M., Cassenote, A., Guerra, A., Guilloux, A. G., Brandão, A. P., Miotto, B. A., Almeida, C. de J., Gomes, J. de O., & Miotto, R. A. (2020). *Demografia médica no Brasil 2020*. Conselho Federal de Medicina.

Scheffer, M., & Dal Poz, M. (2015). The privatization of medical education in Brazil: Trends and challenges. *Human Resources for Health, 13*(1), Article 96. https://doi.org/10.1186/s12960-015-0095-2

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53-55. https://doi.org/10.5116%2Fijme.4dfb.8dfd

Tavakol, M., & Dennick, R. (2013). Psychometric evaluation of a knowledge based examination using Rasch analysis: An illustrative guide: AMEE guide no. 72. *Medical Teacher, 35*(1), e838-e848. https://doi.org/10.3109/0142159X.2012.737488

Tavakol, M., & Dennick, R. (2017). The foundations of measurement and assessment in medical education. *Medical Teacher, 39*(10), 1010-1015. https://doi.org/10.1080/0142159x.2017.1359521

Troncon, L. E. (2019). Exames de licenciamento: Um componente necessário para avaliação externa dos estudantes e egressos dos cursos de graduação em Medicina. *Interface – Comunicação, Saúde, Educação, 24,* Artigo e190576. https://doi.org/10.1590/Interface.190576

Vleuten, C. van der. (2016). Revisiting 'Assessing professional competence: From methods to programmes'. *Medical Education, 50*(9), 885-888. https://doi.org/10.1111/medu.12632

Vleuten, C. van der, Freeman, A., & Collares, C. F. (2018). Progress test utopia. *Perspectives on Medical Education, 7*(2), 136-138. https://doi.org/10.1007/s40037-018-0413-1

Vleuten, C. van der, & Schuwirth, L. W. T. (2005). Assessing professional competence: From methods to programmes. *Medical Education, 39*(3), 309-317. https://doi.org/10.1111/j.1365-2929.2005.02094.x

Vleuten, C. van der, & Schuwirth, L. W. T. (2019). Assessment in the context of problem-based learning. *Advances in Health Sciences Education, 24*(5), 903-914. https://doi.org/10.1007/s10459-019-09909-1

Vleuten, C. van der, Schuwirth, L. W. T., Muijtjens, A. M. M., Thoben, A. J. N. M., Cohen-Schotanus, J., & Boven, C. P. A. van. (2004). Cross institutional collaboration in assessment: A case on progress testing. *Medical Teacher, 26*(8), 719-725. https://doi.org/10.1080/01421590400016464

Walsh, J. L., Harris, B., Denny, P., & Smith, P. (2018). Formative student-authored question bank: Perceptions, question quality and association with summative performance. *Postgraduate Medical Journal, 94*(1108), 97-103. https://doi.org/10.1136/postgradmedj-2017-135018

Wrigley, W., Vleuten, C. van der, Freeman, A., & Muijtjens, A. (2012). A systemic framework for the progress test: Strengths, constraints and issues: AMEE guide no. 71. *Medical Teacher, 34*(9), 683-697. https://doi.org/10.3109/0142159x.2012.704437