RBEM Revista Brasileira de Educação Médica

ORIGINAL ARTICLE

DOI: https://doi.org/10.1590/1981-5271v49.1-2023-0262.ING

Evaluation of the Progress Test of a Medical school according to the SOLO Taxonomy theoretical framework

Avaliação do Teste do Progresso de uma faculdade de Medicina pelos pressupostos da Taxonomia SOLO

Pedro Paulo Trindade Resende[1] pedroresende@ufsj.edu.br
Alexandre de Araújo Pereira[2] alexandre.pereira@unifenas.br
José Maria Peixoto[2] jmpeixoto.prof@gmail.com

ABSTRACT

Introduction: the training of future graduates from medical schools for responsible and qualified health care practice is a significant challenge. The goal is for them to be equipped to solve problems that require higher-order cognitive skills. Thus, evaluating the acquisition of such competencies becomes crucial. One assessment method that has been gaining attention in medical education is the Progress Test (PT). Cognitive theories have advanced educational research related to assessment processes. In our study, we used the Structure of Observing Learning Outcome (SOLO) taxonomy to evaluate and categorize the items of the PT applied at a medical school. The SOLO taxonomy (ST) allows for the necessary cognitive analysis required for performing specific tasks, enabling a comprehensive observation of the student's understanding. We also applied the Classical Test Theory (CTT) in our study, calculating the difficulty index (DFI) and discrimination index (DI) for each multiple-choice item (MCI) of the PT and correlating them with the SOLO classification.

Objective: the objective of this study is to evaluate the characteristics of the PT applied in a private medical school, analyzing its items based on the assumptions of ST and correlating them with CTT.

Materials and Methods: this is a descriptive study with a quantitative and qualitative approach. According to the principles of ST, we conducted the analysis and characterization of the items from a PT applied in a private medical school and correlated them with the DFI and DI.

Results: we found a balance between surface learning (SL) and deep learning (DL) across the total items, as well as a direct relationship between the levels of DL and MCIs consisting of clinical cases. We did not find statistically significant differences between the SOLO categories regarding the means of DFI and DI.

Conclusion: the analysis of assessment activities should not be restricted to psychometric properties. Taxonomic tools, such as the ST, can significantly aid in conducting these activities, aligning assessments with the curriculum and facilitating the creation of tests appropriate for the desired level of learning, thereby promoting effective teaching progression.

Keywords: Taxonomy; Education, Medical; Academic Performance; Educational Measurement.

RESUMO

Introdução: *A capacitação de futuros egressos das faculdades de Medicina para uma prática assistencial responsável e qualificada de atenção à saúde é um desafio. O objetivo é que estejam aptos para a resolução de problemas que demandem habilidades cognitivas de ordem superior. Sendo assim, a avaliação da aquisição de tais competências se torna muito importante. Um método de avaliação que vem ganhando atenção no ensino médico é o Teste do Progresso (TP). Teorias cognitivas têm trazido progresso em pesquisas educacionais relativas aos processos de avaliação. Utilizamos no nosso estudo a Taxonomia Structure of Observing Learning Outcome (SOLO) para avaliar e categorizar os itens do TP aplicado em uma faculdade de Medicina. A Taxonomia SOLO (TS) possibilita a análise cognitiva necessária para a realização de determinadas tarefas, permitindo uma observação integral da compreensão do entendimento do aluno. Utilizamos também no nosso estudo a Teoria Clássica dos Testes (TCT) e calculamos para cada item de múltipla escolha (IME) do TP o índice de dificuldade (IDF) e o índice de discriminação (ID), e os correlacionamos com a classificação SOLO.*

Objetivo: *Este estudo teve como objetivo avaliar características do TP aplicado em uma faculdade privada de Medicina, analisando seus itens pelos pressupostos da TS e correlacionando-os com a TCT.*

Método: *Trata-se de um estudo descritivo, de abordagem quantitativa e qualitativa. De acordo com os princípios da TS, foram realizadas a análise e a caracterização dos itens da prova de um TP aplicado em uma faculdade privada de Medicina e a correlação com o IDF e o ID.*

Resultado: *Verificamos um equilíbrio entre a aprendizagem superficial (AS) e a aprendizagem profunda (AP) no total de itens e uma relação direta entre os níveis de AP e IME compostos por casos clínicos. Não verificamos diferença estatisticamente significativa entre as categorias SOLO quanto às médias do IDF e do ID.*

Conclusão: *A análise das atividades avaliativas não deve ser restringida às propriedades psicométricas. Ferramentas taxonômicas, como a TS, podem auxiliar de maneira significativa a realização dessas atividades, de modo a conciliar as avaliações ao currículo, possibilitar a realização de provas adequadas ao nível de aprendizagem desejável e favorecer a progressividade do ensino.*

Palavras-chave: *Taxonomia; Educação Médica; Desempenho acadêmico; Avaliação Educacional.*

[1] Universidade Federal de São João Del Rei, São João Del Rei, Minas Gerais, Brazil.
[2] Universidade José do Rosário Vellano, Belo Horizonte, Minas Gerais, Brazil.

Chief Editor: Rosiane Viana Zuza Diniz.
Associate Editor: Daniela Chiesa.

Received on 10/09/23; Accepted on 11/17/24.

Evaluated by double blind review process.

DOI: https://doi.org/10.1590/1981-5271v49.1-2023-0262.ING

INTRODUCTION

The purpose of medical schools is to prepare future graduates for responsible and qualified care practice. In this sense, the learning assessment process is essential, as it makes it possible to observe the results of educational interventions[1]. The evaluation procedure is an important tool in the teaching arsenal, being an instrument that drives learning [2-4]. However, the evaluation methods used are not usually designed to identify the evolution of students' cognitive abilities; in general, they evaluate the acquisition of the studied contents, often requiring simple memorization.

To contribute to the learning process, the assessment must be planned in such a way as to reflect the objectives expected to be achieved, at the cognitive, affective and psychomotor levels, and allow both teachers and students to identify their progress and points for improvement[5].

An evaluation method that has been gaining attention in medical schools is the Progress Test (PT), which evaluates the students' longitudinal cognitive performance[6]. The PT is applied every six months or annually to all students of the course, simultaneously, and consists of multiple choice items (MCIs). The PT content comprises the entire curricular matrix, being related to its final objectives and based on the National Curriculum Guidelines for the Undergraduate Medical Course [7,8].

The PT must demand from the students much more than memorization. The MCIs are prepared by teachers from the institution itself or from consortia created for this purpose, and a gradual cognitive gain is expected over the semesters[8].

Due to the importance that the PT has been gaining in medical education, it is relevant to evaluate whether the quality of its items meets the premises of a good evaluation process. This includes verifying that the PT MCIs have:

- **Discriminative ability:** items should be able to differentiate between students with different levels of knowledge and abilities.
- **Quantitative balance regarding the difficulty of the items:** There should be a balanced distribution of items with varying degrees of difficulty, ensuring that the test is challenging enough for all students, without being overly easy or difficult.
- **Appropriate cognitive complexity**: Items should require not only memorization, but the ability to make correlations, inferences, and generalizations, reflecting the cognitive demands of medical practice.

It is important to differentiate between difficulty and complexity regarding evaluative items. Difficulty is related to factors that offer obstacles to solving a problem, while complexity involves the cognitive skills needed to solve it. Difficulty is usually evaluated statistically, while complexity is evaluated qualitatively, frequently using educational taxonomies[9,10]. Therefore, an item may have a high level of difficulty and low complexity, or vice versa.

Among the statistical methods for psychometric evaluation, the Classical Test Theory (CTT) allows the evaluation of the difficulty index (DFI) and the discrimination index (DI) of the items that comprise the test. The DFI is found by calculating the ratio between the number of students who answered correctly and the total number of students submitted to the item. This index ranges from 0 to 1. Table 1 shows a classification of the items of a test in relation to the DFI and the expected percentage of distribution in an evaluation[11].

The DI is calculated by the difference between the percentage of correct answers to a given item of the students who performed better on the test and those who had a worse performance. To attain this calculation, candidates will be allocated into three groups: the upper group (27% of the highest scores), the lower group (27% of the lowest scores) and the intermediate group, with the remaining 46% of the candidates[11]. Table 2 shows the criteria for the DI values and the item classification according to its power of discrimination.

Regarding the item complexity, they can be evaluated through educational taxonomies, which are classification

Table 1. Criteria for distribution and classification of the item difficulty degree by CTT.

Optimal quantity of items in an assessment (expected %)	Item difficulty index	Item classification regarding the Difficulty Index
10%	Greater than 0.9	Very easy
20%	From 0.7 to 0.9	Easy
40%	From 0.3 to 0.7	Median
20%	From 0.1 to 0.3	Difficult
10%	Up to 0.1	Very difficult

Source: Vilarinho, 2015, p. 27.

Table 2. Values of the power of discrimination and item classification according to the CTT.

Values	Classification
Discrimination < 0.20	Deficient item, should be rejected
0.20 ≤ Discrimination < 0.30	Marginal item, subject to re-creation
0.30 ≤ Discrimination < 0.40	Good item, but subject to improvement
Discrimination ≥ 0.40	Good item

Source: Vilarinho, 2015, p 28.

DOI: https://doi.org/10.1590/1981-5271v49.1-2023-0262.ING

systems that allow the categorization of the learning levels achieved by the students, being useful for the evaluation system and planning of educational goals[12]. Among the existing educational taxonomies, the SOLO (Structure of Observed Learning Outcomes) Taxonomy was conceived under the idea that individuals learn different contents in ascending stages of complexity[13,14].

Biggs and Collis[15] proposed a categorization of the stages of content understanding, called "modes of thought", based on the Piagetian stages. This system identifies different degrees of thought formalization, allowing the quality of learning to be assessed.

The SOLO Taxonomy (ST), derived from these concepts, classifies the structure of demonstrated learning into five progressive levels of cognitive complexity[15], as shown in Figure 1:

Pre-structural (SOLO 1): inadequate responses, irrelevant or incoherent information.

Unistructural (SOLO 2): responses that are directed to a single element of the task and therefore inconsistent.

Multistructural (SOLO 3): responses identify more than one element of the task, but there is no integration of the information, leading to inconsistencies.

Relational (SOLO 4): various information is identified and relationships are established in a coherent way, with an understanding of the whole, with no inconsistencies.

Extended abstract (SOLO 5): the response goes beyond the elements of the item, moving towards abstraction and generalization.

Studies have identified two main forms of learning: one called surface and the other deep learning. Surface learning (SL) is characterized by the reproduction of content without connections or reflections, while deep learning (DL) involves an intrinsic and reflective understanding, requiring sophisticated cognitive processes[15]. SL is formed by the SOLO 2 and SOLO 3 levels and the DL, by the SOLO 4 and SOLO 5 levels. SL is based on the retention of concrete details, through memorization. DL is more complex, requiring information relations, qualifying the individual to understand mechanisms and principles and to make generalizations or theorizations [15,17].

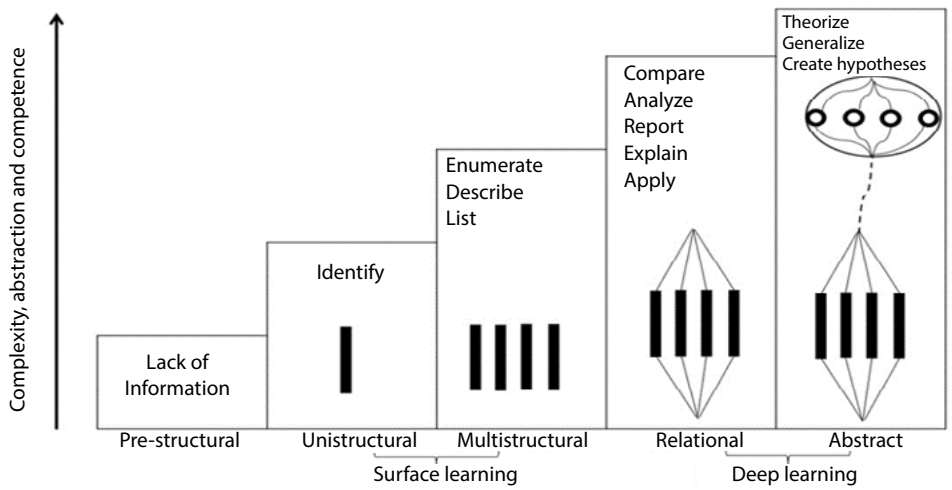
Throughout the training, the students' progress occurs upwards, from a concrete to an abstract understanding, developing skills to establish relationships and make inferences, reflecting an increase in the ability to handle information consistently and make generalizations. This ascending evolution of the students' cognitive process can be categorized as learning cycles, which represent the way students understand and operate the studied content, from the most concrete to the most abstract one[18].

Considering the exposed reasons, the objective of this study was to analyze the psychometric characteristics of a PT assessment by CTT, in relation to its difficulty and discrimination indexes, in addition to categorizing the cognitive complexity of its items by the ST assumptions. The ST was selected for this study due to its ability to classify the complexity of the learning structure demonstrated by students in a specific task, identifying the thought processes involved, and the possibility of differentiating surface learning from deep learning[19,20].

MATERIALS AND METHODS

This is a descriptive study, with a quantitative and qualitative approach. We analyzed the PT applied in the first semester of 2022 in the Medicine Course at Universidade Prof. Edson Antônio Velano, in the Belo Horizonte Campus (TPU2022-1), applied to all students, from the 1st to the 12th semesters, simultaneously, containing 120 MCIs, with the content being divided into the following areas of knowledge: surgery, internal

Figure 1. SOLO taxonomy.



Source: Translated from Biggs and Tang[16].

DOI: https://doi.org/10.1590/1981-5271v49.1-2023-0262.ING

Chart 1. Categorization system for TPU2022-1 items.

Question SOLO Category	Examples of Command Verbs and Their Relation with SOLO Categories	Knowledge addressed in the item		Cognitive procedures involved for item resolution
		Number of Topics used	Relationship between topics	
Abstract (SOLO 5)	Discuss, hypothesize, evaluate, reason, estimate, criticize, interpret, predict, reflect, program, judge, generalize, implement	Two or more topics	Requires association between topics	induction and/or deduction; requires identification of relevant information not commonly discussed in medical school, creation of hypotheses and generalizations
Relational (SOLO 4)	Explain, integrate, refer, analyze, compare, interpret, build, plan, summarize, relate, argue			induction and/or deduction; requires identification of relevant information frequently discussed in the medical course
Multistructural (SOLO 3)	Describe, execute, solve, apply, combine, complete, classify, enumerate		Used alone	induction and/or deduction; requires identification of relevant information frequently discussed in the medical course
Unistructural (SOLO 2)	Identify, decide, organize, reproduce, choose, find, recognize, tell, search, paraphrase	A single topic	Not applicable	induction and/or deduction; requires identification of relevant information frequently discussed in the medical course

Source: Adapted from Ceia[18] and Pereira[19].

medicine, gynecology-obstetrics, pediatrics and collective health, with 24 items for each area.

The TPU2022-1 test was based on the TEP MINAS 2019 test matrix, which was prepared by the team of the Minas Gerais Consortium of Medical Schools for the Progress Test (TEP MINAS 1). TEP MINAS 1 comprise eight medical schools in the state of Minas Gerais, including public and private entities.

The orders (guide for the creation of the questions) were sent to the teachers of the institutions according to their area of expertise. When delivered, the questions were reviewed, and the necessary corrections were made.

The TPU2022-1 MCIs were evaluated using CTT and categorized according to the ST criteria. The categorization of MCIs by ST was performed through an adaptation of the model used for classifying the item complexity of an assessment proposed by Mário Ceia[21]. According to this model, the item of an evaluation is analyzed based on the expected answer to the question, considering three parameters: amount of knowledge necessary for its resolution, cognitive operations involved in solving the problem, and complexity of the requested answer. Chart 1 shows the categorization system proposed for this study, which was adapted from the studies by Ceia[21] and Pereira[10].

Based on the information in Chart 1, the Item Categorization Form (ICF), Chart 2, was created, where, for each item of TPU2022-1, the statement was transcribed, followed by the alternatives and the answer key of the question. There are also spaces in the ICF for analyzing the content of the item, the cognitive procedures necessary for its resolution and, finally, a space for its categorization according to ST. A list of verbs frequently associated with each SOLO category was attached to the ICF so that evaluators could consult and better adjust their opinions.

Chart 2. Item Categorization Form.

Below you will find the transcript of question n. ..... of the UNIFENAS-BH Progress Test of the year 20..... Read the question statement and evaluate the answer marked as correct. Starting from the correct answer, analyze the contents and cognitive procedures required to solve the question, considering the Item Categorization System presented to you. At the end, classify the complexity of the question according to the SOLO Taxonomy.
Statement of the item: Alternatives: Answer key: .....
Content analysis:
Analysis of procedures:
SOLO Category of the question: ( ) SOLO 2 ( ) SOLO 3 ( ) SOLO 4 ( ) SOLO 5

Source: Prepared by the authors

Two medical professors, PhDs, who received training on the ST assumptions for MCI categorization, participated in the categorization of MCIs, in addition to the main researcher. A total of 33.33% of the TPU2022-1 items were selected, which correspond to 40 items, using a systematic probabilistic sampling method, so that the three evaluators could perform their analyses to identify any categorization bias. Each evaluator received the selected items and, after an individual analysis, filled out the ICF. The evaluators' analyses were compared and

DOI: https://doi.org/10.1590/1981-5271v49.1-2023-0262.ING

a high agreement rate, of 95%, was verified. Adjustments were made by consensus, allowing a categorization calibration by the main researcher. The other items were analyzed and classified only by the main researcher.

All items of the TPU2022-1 were classified according to the ST principles, in four levels of cognitive complexity: unistructural, multistructural, relational and extended abstract. The pre-structural level was not included, since the purpose was to analyze the MCIs of the PT in relation to the cognitive complexity required for their resolution, so items with this categorization were not expected. After the categorization, the items were subdivided into two learning categories: surface (SOLO 2 and 3) and deep learning (SOLO 4 and 5).

The MCIs were also submitted to psychometric analysis by CTT, where the discrimination index (DFI) and the difficulty index (DI) were calculated for each item. As the PT is applied to students with different learning cycles, the tests of the students attending the last year of the course were considered as the reference for the analysis by the CTT, since it is a representative sample of students who completed 83.33% of the curricular matrix.

Aiming to investigate whether there was a significant difference in the means of the measurements of the DFI and DI parameters regarding the SOLO levels, the Analysis of Variance with 1 factor (One-Way) was applied to the data. Levene's Test for Equality of Variance was used to investigate whether the variances between the categories were statistically different. In addition, the effect size (Partial ETA squared) was calculated. To investigate whether there was a significant difference in the means of the measurements of the DFI and DI parameters regarding the SOLO categories that represent surface learning and deep learning (2 or 3 × 4), the Student's *t* test for Independent Samples was applied to the data. Levene's Test for Equality of Variance was again used to investigate whether the variances between the categories were statistically different. Moreover, the Effect Size (Cohen's *d)* was calculated. The results were considered significant for a probability of significance of less than 5%, with at least 95% confidence in the presented conclusions.

The present study was approved by the Research Ethics Committee of UNIFENAS, under CAAE number 56009222.9.0000.5143, Opinion number 5.379.183; the waiver of the Informed Consent Form was requested, and the Data Use Commitment Term and the Assent Form were sent.

RESULTS

A balance was found between surface and deep learning in the TPU2022-1 items. Approximately 41% were classified as SL and 59.2% as DL, as shown in Table 3. When analyzed by specific areas of knowledge, this balance was identified in the areas of surgery, gynecology-obstetrics, and pediatrics. We did not find a balance between surface and deep learning in the areas of internal medicine and collective health,. In internal medicine, we found that 87.5% of the analyzed items were related to DL and 12.5% to SL. In the area of collective health, we found that about 83.3% of the items were related to SL and 16.7% to DL. Table 3 divides the TPU2022-1 items between SL and DL.

We did not identify any item related to the extended abstract level, which is the level of greater cognitive complexity of the ST. Table 4 shows the results of the analysis, according to the ST levels of complexity, by areas of knowledge: surgery, internal medicine, gynecology-obstetrics, pediatrics, and collective health.

Table 3. Distribution of TPU2022-1 items between surface and deep learning.

Area/Soil	Surface Learning	Deep Learning
Surgery	7 (29.2%)	17 (70.8%)
Internal Medicine	3 (12.5%)	21 (87.5%)
Gynecology/Obstetrics	10 (41.7%)	14 (58.3%)
Pediatrics	9 (37.5%)	15 (62.5%)
Collective Health	20 (83.3%)	4 (16.7%)
Total	49 (40.8%)	71 (59.2%)

Source: Prepared by the authors (2023).

Table 4. Distribution of TPU2022-1 items by ST levels.

Area/SOLO	Unistructural	Multistructural	Relational	Abstract
Surgery	3 (12.5%)	4 (16.7%)	17 (70.8%)	0
Internal Medicine	2 (8.3%)	1 (4.2%)	21 (87.5%)	0
Gynecology/Obstetrics	7 (29.2%)	3 (12.5%)	14 (58.3%)	0
Pediatrics	7 (29.2%)	2 (8.3%)	15 (62.5%)	0
Collective Health	18 (75.0%)	2 (8.3%)	4 (16.7%)	0
Total	37 (30.8%)	12 (10%)	71 (59.1%)	0

Source: Prepared by the authors, research data (2023).

DOI: https://doi.org/10.1590/1981-5271v49.1-2023-0262.ING

We found that the items consisting of clinical cases that required problem-solving skills provided greater exploration of DL. Table 5 shows the strategy used in the formulation of the items, based on the presence or absence of clinical cases. Most of the items (84.2%) had the presence of a clinical case. The need for a list of topics for the resolution of clinical cases is verified in most of the items in the areas of knowledge, with the exception of the collective health area, in which a considerable portion of the items (54.2%) did not contain clinical cases, and when a clinical case was present, which occurred in 11 items (45.8%), only in two was the list of topics required for its resolution.

Regarding the psychometric analysis, we found that 10.8% of the items in TPU2022-1 had an DFI at the easy level, 50.9% at the median level, and 25% at the difficult level. The proportion of very difficult and very easy items was 13.3%. Table 6 shows the results according to the DFI.

In our analysis, we found that 79.2% of the items had a DI lower than 20%. About 2% had a DI greater than 40%. Table 7 shows the results according to the DI.

Table 8 shows that no statistically significant difference was identified between the three SOLO categories regarding the CTT parameters. It should be noted that the Effect Size is considered small, therefore, corroborating the statistical non-significance of the test. It is emphasized that no statistically significant difference was observed between the variances (*Levene*→ $p > 0.05$). Therefore, there is no need to apply the Welch's test.

Table 9 shows that there was no statistically significant difference between the SOLO categories that represent surface and deep learning and the CTT parameters.

At the end of this article, there is a link to access the analysis of all TPU2022-1 items according to the assumptions of the ST.

Table 5. Presence or absence of clinical cases in the TPU2022-1 items.

SOLO Area/Item	Absence of clinical case	Presence of clinical case
Surgery	1	23
Internal Medicine	0	24
Gynecology/Obstetrics	3	21
Pediatrics	2	22
Collective Health	13	11
Total	19	101

Source: Prepared by the authors, research data (2023).

Table 6. TPU2022-1 Difficulty Index.

% of correct answers	Number of questions	Frequency (%)
≤ 10.0% (Very Easy)	1	0.8
From 10.1 to 30.0% (Easy)	13	10.8
From 30.1 to 70.0% (median)	61	50.9
From 70.1 to 90.0% (Hard)	30	25.0
> 90.0% (Very difficult)	15	12.5
Total	120	100.0

Source: Prepared by the authors, research data (2023).

Table 7. TPU2022-1 Discrimination Index.

Discrimination	Number of questions	Frequency (%)
< 20	95	79.1
0.20 ≤ to < 0.30	15	12.5
0.30 ≤ to < 0.40	8	6.7
≥ 40	2	1.7
Total	120	100.0

Source: Prepared by the authors, research data (2023).

DISCUSSION

The objective of this study was to evaluate the PT items applied in a private educational institution, in relation to their psychometric characteristics and cognitive complexity through ST. Our results found a balance in TPU2022-1

Table 8. Comparative analyses between the SOLO categories regarding the CTT evaluation parameters (difficulty and discrimination).

SOLO Category	Descriptive measures		p-value
	N	Mean ± SD	
Difficulty Index			
SOLO 2	37	60.9 ± 22.8	0,821 F2, 117 = 0.197
SOLO 3	12	62.1 ± 19.4	
SOLO 4	71	58.5 ± 24.4	
Overall	120	59.6 ± 23.3	
Discrimination Index			
SOLO 2	37	15.0 ± 10.6	0.484 F2, 117 = 0.731
SOLO 3	12	11.3 ± 7.2	
SOLO 4	71	13.4 ± 9.6	
Overall	120	13.7 ± 9.7	

Database: 120 questions (SOLO 2 →37 cases, SOLO 3→12 cases and SOLO 4 →71 cases)
Note: **SD** → Standard deviation; ***P*:** Probability of significance of the Analysis of Variance with 1 factor (One-Way).
F: Statistics of Analysis of Variance with 1 factor (One-Way).
- **Levene's test for equality of variance:** $p > 0.05$. For both variables.
- **Effect size** (η^2): $\eta^2 \leq 0{,}01$ → Effect size: Small (both variables).

DOI: https://doi.org/10.1590/1981-5271v49.1-2023-0262.ING

Table 9. Comparative analysis between the SOLO categories (surface and deep learning) regarding the CTT evaluation parameters (difficulty and discrimination).

SOLO Learning Category	Descriptive measures		p-value
	N	Mean ± SD	
Difficulty Index			
Surface	49	61.2 ± 21.8	0.544
Deep	71	58.5 ± 24.4	*t118 = 0.609*
Overall	120	59.6 ± 23.3	
Discrimination Index			
Surface	49	14.1 ± 9.9	0.684
Deep	71	13.4 ± 9.6	*t118 = 0.408*
Overall	120	13.7 ± 9.7	

Database: 120 questions (SOLO 2 →37 cases, SOLO 3→12 cases and SOLO 4 →71 cases)
Note: **SD** →Standard deviation; ***P*:** Probability *of* significance of the Analysis of Variance with 1 factor (One-Way).
F: Statistics of Analysis of Variance with 1 factor (One-Way).
- **Levene's test for equality of variance:** p > 0.05. For both variables.
- **Effect size (*Cohen's d*):** $d \leq 0.12$ →Effect size: Small (both variables).

between the frequency of items related to SL and DL, with a predominance of items related to DL. The areas of surgery, gynecology-obstetrics and pediatrics were the ones that showed this balance the most.

Researchers state that a balance in the item distribution of an assessment, related to the levels of cognitive complexity, contributes to a better assessment of learning. Therefore, a balanced distribution according to learning taxonomies attains great importance[22]. Despite the importance attributed to this balance, it is assumed that medical course graduates are able to solve complex problems. The expectation is that students in the last semesters be able to solve activities, such as problem solving and decision-making[23]. Therefore, we believe that activities aimed at evaluating DL should prevail in the PT. The ST helps in the construction and selection of appropriate assessment items to verify the acquisition of attributes expected of a graduate and that foster analysis rather than simple memorization[24].

The items classified as DL of the TPU2022-1 contained, in most cases, a clinical case and the statement required problem-solving skills and knowledge integration, requiring clinical reasoning. Clinical reasoning requires a knowledge base, enabling students to generate hypotheses, establish diagnoses, and offer a conduct for solving clinical problems[25,26]. The use of genuine clinical problem solving is an efficient measure of clinical reasoning analysis[27].

The construction of higher-order thinking is crucial in the training and practice of medicine[28]. DL is associated with better effectiveness in medical education and is more related to the students' ability to keep updated after their training[17].

Another important applicability of learning taxonomies, such as ST, is to offer students data on their level of cognitive thinking. This becomes very significant at this time of paradigm shift in Higher Education, in which autonomous learning, centered on the student, is recommended[29].

The non-identification of any TPU2022-1 item related to the extended abstract level may be related to the use of MCIs, which may bring a certain limitation to the evaluation of the higher taxonomic level[30,31]. Other evaluation strategies, such as discursive items, are adequate to demonstrate this level. However, it would be unfeasible to carry out an exam with this type of item, in which the objective is to evaluate a significant number of students, with a very extensive content[32]. Thus, it is necessary to discuss a methodology that provides the construction of items at a SOLO 5 level of the ST, through MCIs.

The multiple-choice question method is widely used in the evaluation processes of medical schools. MCI-based assessments, when well designed, have the ability to assess students at higher levels of knowledge, making this task a challenging one[33]. The introduction of clinical cases in multiple-choice tests improves the quality of this evaluation process to measure clinical reasoning[34]. That was observed in TPU2022-1, where many items required more complex reasoning through clinical case resolution.

The psychometric analysis of the MCIs of an assessment is very important to confirm its quality. The MCIs should be evaluated to verify their validity and reliability[35]. It is very relevant to assess the reason for choosing one answer option to the detriment of others and the reason why the most answered alternative is not the correct one[36].

We found a satisfactory result related to the DFI, with a lower percentage of very easy and very difficult items (13.3%), with most items (86.7%) being within an acceptable range for an evaluation process. Most of the MCIs were at an unsatisfactory level in relation to the DI, not allowing a differentiation between the students with the best and worst performances, suggesting the review of a significant number of items. This can raise questions about interpretations and conclusions, based on a test with this particularity.

Some factors can affect the DI of an item, such as question ambivalence, excessively difficult or easy solution, presence of topics not discussed in the curriculum, students' poor preparation, students' motivation, number of questions, time to resolution, and environmental factors, such as temperature, noise, and ventilation[37].

No correlation was identified between the CTT parameters (DFI and DI) and the ST categories. There was also no correlation between the SOLO categories, which represent surface learning and deep learning, and the CTT parameters.

DOI: https://doi.org/10.1590/1981-5271v49.1-2023-0262.ING

Bicudo et al.[38] demonstrated, in a progress test applied to ten Brazilian medical schools, that items related to high taxonomic levels achieved a better performance in discrimination indexes.

When analyzing the questions through the ST, it is verified that the degree of complexity is not a determinant of their level of difficulty[30,39]. The degree of difficulty of the questions is assessed by statistical methods, whereas the level of complexity can be established by qualitative methodologies, such as ST[10]. This information demonstrates that the quality of an assessment must take into account, in addition to psychometric data, the cognitive abilities involved in solving the items, that is, the level of cognitive complexity required, data that are not evaluated by the commonly used psychometric methodologies.

For a better performance of evaluation processes, such as PT, it is very important to prepare the faculty for the performance of its items. This requires a continuing education program for teachers in medical schools for the development of quality assessment items[40]. Learning taxonomies can contribute considerably to the creation of well-planned assessments, contributing to an effective learning environment[24].

There are many evaluation techniques. No single method should be used in the evaluation processes of medical schools[41]. An association of techniques is indispensable to attain a satisfactory proof of the students' performance[42,43].

There are few studies in the literature that analyze evaluations external to the ST. Mol[44] reviewed Brazilian studies using the ST and found 14 studies, of which ten are articles and four are dissertations. No theses were found that addressed the ST. Only two studies were related to higher education and none in the area of medical education. Ferreira and Rocha[45] conducted a survey of the theses and dissertations defended in Brazil that used the ST. They found twelve studies, most in the area of exact sciences and none related to medical education.

Some limitations of the present study should be considered. The first is related to the fact that the study comprised only one PT exam, from a specific medical school. The analysis of a larger amount of evidence, including other medical schools, may provide greater support for a more adequate interpretation of the data. The fact that the categorization of all items was performed by a single evaluator may also be a limiting factor; however, it is important to observe the attempt to standardize this analysis by three evaluators with a high rate of agreement, a fact that favors the applicability of this analysis methodology in the daily life of a school, where the teacher often does not have a team of evaluators available.

Nevertheless, this study presents relevant contributions to the process of reflection and development of evaluations in medical education. It demonstrates the need to balance analysis methodologies beyond psychometrics, including those that consider the cognitive processes involved in solving the questions. It demonstrates that the inclusion of clinical cases favors the analysis of deep learning, and the need to seek strategies for the creation of items with good discriminatory capacity and balance in terms of difficulty should be considered. The study also presents a methodology for categorizing the cognitive complexity of MCIs that should be tested in new studies.

The data of this study can contribute to the inclusion of an analysis group, which considers, in addition to the statistical data of psychometrics, its characteristics in terms of cognitive complexity.

CONCLUSION

The analysis of TPU2022-1 allowed us to investigate characteristics of this tool, useful in the evaluation of the students' longitudinal cognitive performance. We found a balance between SL and DL when all items were analyzed. However, we did not observe this balance in the areas of internal medicine and collective health. We found a direct relationship between the DL levels of the ST and MCIs consisting of clinical cases. We found a good relationship for the DFI, but not for the DI, suggesting the need for a critical analysis of the items. We found no statistically significant differences between the SOLO categories that represent SL and DL regarding the means of the DFI and the DI, indicating that these methodologies evaluate different particularities of the MCIs.

The PT is a valuable instrument for evaluating teaching and should be encouraged in medical schools. Taxonomic tools, such as the ST, can significantly help in the performance of evaluation activities, reconciling evaluations with the curriculum, allowing the performance of tests that are appropriate to the desirable level of learning, favoring teaching progressiveness. The qualification of the faculty to carry out evaluation activities is necessary. Workshops for assistance in the creation of items should be offered to the teachers.

Complementary studies are essential to increase the consistency of the use of this taxonomic tool in medical education.

AUTHORS' CONTRIBUTIONS

Pedro Paulo Trindade Resende participated in the creation of the research project, literature review, data analysis, discussion of the results and in the writing and review of the manuscript. Alexandre de Araújo Pereira participated in the creation of the research project, data analysis, discussion of the results and in the writing and review of the manuscript. José Maria Peixoto

DOI: https://doi.org/10.1590/1981-5271v49.1-2023-0262.ING

participated in the creation of the research project, discussion of the results and the writing and review of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

SOURCES OF FUNDING

The authors declare no sources of funding.

TPUBH2022-1 ITEMS ACCESS LINK

https://docs.google.com/document/d/1V3SM_cLmPYgfWnprkZfRw16QK0C-WjxR/edit?usp=drive_link&ouid=103188252538609047663&rtpof=true&sd=true

REFERENCES

1. Champlain AFC. Setting and maintaining standards in multiple-choice examinations: guide supplement 37.2 – Viewpoint. Med Teach. 2010;32:436-7.
2. O'Shaughnessy SM, Joyce P. Summative and formative assessment in medicine: the experience of an anaesthesia trainee. Int J Higher Educ. 2015;4(2):198-206.
3. Ferris H, O'Flynn, D. Assessment in medical education: what are we trying to achieve? Int J Higher Educ. 2015;4(2):139-44.
4. Prashanti E, Ramnarayan K. Ten maxims of formative assessment. Adv Physiol Educ. 2019;43:99-102.
5. Hadie SNH. The application of learning taxonomy in anatomy assessment in medical school. Education in Medicine Journal. 2018;10(1):13-23.
6. Reberti AG, Monfredini NH, Ferreira Filho OF, Andrade DF, Pinheiro CEA, Silva, JC. Teste de Progresso na escola médica: uma revisão sistemática acerca da literatura. Rev Bras Educ Med. 2020;44(1):1-9.
7. Pinheiro OL, Spadella MA, Moreira HM, Ribeiro ZMT, Guimarães APC, Almeida Filho OM, et al. Teste de Progresso: uma ferramenta avaliativa para a gestão acadêmica. Rev Bras Educ Med. 2015;39(1):68-78.
8. Baldim YL, Vicente C A O, Arcuri MB. O teste de progresso sob a visão do discente. Revista da Faculdade de Medicina de Teresópolis. 2018;2(1):41-54.
9. Condé FN. Análise empírica de itens. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais; 2001.
10. Pereira VCAS. Aplicação da Taxonomia SOLO na análise da qualidade da avaliação: validação do método analítico por aplicação aos exames nacionais de Matemática A entre 2006 e 2014. Covilhã; 2019.
11. Vilarinho APL. Uma proposta de análise de desempenho dos estudantes e de valorização da primeira fase da OBMEP [dissertação]. Brasília: Universidade de Brasília; 2015.
12. Aripin MA, Hamzah R, Setya P, Hisham MHM, Mohd Ishar MI. Unveiling a new taxonomy in education field. International Journal of Evaluation and Research in Education. 2020;9(3):524-30.
13. Amantes A, Oliveira E. A construção e o uso de sistemas de categorias para avaliar o entendimento dos estudantes. Revista Ensaio. 2012;24(2):61-79.
14. Biggs J. Study Process Questionnaire Manual. Student. Approaches to Learning and Studying. Australian Council for Educational Research; 1987.
15. Biggs J, Collis K. Evaluating the quality of learning: the SOLO Taxonomy. New York: Academic Press;1982.
16. Biggs, J. Calidade del Aprendizage Universitário. Narcea; 2006.
17. Rossi GZ, Fischer JMS, Rocha SR, Casalecchi GA, Avó LSR, Germano CMR. Abordagens de aprendizado e sua correlação com ambiente educacional e características individuais em escola médica. Rev Bras Educ Med. 2021;45(3):1-11.
18. Filipe MAER. A Taxonomia SOLO nos Exames Nacionais de Matemática – 9º ano [dissertação]. Lisboa: Universidade Nova de Lisboa; 2011.
19. Yurtyapan MI, Yilmaz GK. An investigation of the geometric thinking levels of middle school mathematics preservice teachers according to SOLO Taxonomy: "Social Distance Problems". Participatory Educational Research. 2021;8(3):188-209.
20. Jimoyiannis A. Using SOLO taxonomy to explore students' mental models of the programming variable and the assignment statement. Themes in Science & Technology Education. 2011;4(2):53-74.
21. Ceia M. A taxonomia SOLO e os níveis de Van Hiele. Encontro de Investigação em Educação Matemática. Coimbra: Sociedade Portuguesa de Ciências da Educação; 2002.
22. Korkmaz F, Unsal S. Analysis of attainments and evaluation questions in sociology curriculum according to the SOLO Taxonomy. Eurasian Journal of Educational Research. 2017;69:75-92.
23. Soobard R, Rannikmäe M. Examining Curriculum Related Progress Using a Context-Based Test Instrument – a comparison of Estonian Grade 10 and 11 Students. Science Education International. 2015;26(3):263-83.
24. Hadie SNH. The application of learning taxonomy in anatomy assessment in medical school. Education in Medicine Journal. 2018;10(1):13-23
25. Peixoto JM, Santos SME, Faria RMD, Moura AS. Processos de desenvolvimento do raciocínio clínico em estudantes de Medicina. Rev Bras Educ Med. 2018;42(1):75-83.
26. Thampy H, Willert E, Ramani S. assessing clinical reasoning: targeting the higher levels of the pyramid. J Gen Intern Med. 2019;34(8):1631-6.
27. Van Der Vleuten CPM, Schuwirth LWT. Assessment in the context of problem-based learning. Adv Health Sci Educ. 2019;24:903-16.
28. Aragão JCS, Almeida LS. Raciocínio clínico e pensamento crítico: desenvolvimento na educação médica. Rev Estud Investig Psicol Educ. 2017;(12):12-20.
29. Jaiswal P. Using constructive alignment to foster teaching learning processes. English Language Teaching. 2019;12(6):10-23.
30. Scully D. Constructing multiple-choice items to measure higher-order thinking. Practical Assessment, Research, and Evaluation. 2017;17(4):1-12.
31. Sprecher EA. Back to the chalkboard: lessons in scaffolding using SOLO taxonomy from school teachers for university educators. Psychology Teaching Review. 2019;25(2):95-102.
32. Kim M, Patel RA, Uchizono JA, Beck L. Incorporation of Bloom's Taxonomy into multiple-choice examination questions for a pharmacotherapeutics course. Am J Pharma Educ. 2012;76(6):1-8.
33. Vegi VAK, Sudhakar PV, Bhimarasetty DM, Pamarth K, Edara L, Kutikuppala LVS, et al. Multiple-choice questions in assessment: perceptions of medical students from low-resource setting. J Educ Health Promot. 2022;11:1-6
34. Modi JN, Anshu, Gupta P, Singh T. Teaching and assessing clinical reasoning skills. Indian Pediatr. 2015;52:787-94.
35. Gajjar S, Sharma R, Kumar P, Rana M. Item and test analysis to identify quality Multiple Choice Questions (MCQs) from an Assessment of Medical Students of Ahmedabad, Gujarat. Indian Journal of Community Medicine. 2014;39(1):17-20.
36. Piton-Gonçalves J, Almeida AM. Análise da dificuldade e da discriminação de itens de Matemática do Enem. Revista Eletrônica de Matemática. 2018;4(2):38-53.
37. Bhattacherjee S, Mukherjee A, Bhandari K, Rout AJ. Evaluation of Multiple-Choice Questions by Item Analysis, from an Online Internal Assessment of 6th Semester Medical Students in a Rural Medical College, West Bengal. Indian Journal of Community Medicine. 2022;47(1):92-5.
38. 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):33-9.
39. Hattie JAC, Brown GTL. Cognitive processes in asTTle: the SOLO taxonomy. AsTTle Technical Report. Wellington: Ministry of Education; 2004.

DOI: https://doi.org/10.1590/1981-5271v49.1-2023-0262.ING

40. Vanderbilt AA, Feldman M, Wood IK. Assessment in undergraduate medical education: a review of course exams. Med Educ Online. 2013;18(1):1-5.

41. Shah SSH, Munir TA, Sabir M, Tipu SA. Psychometric analysis of MCQs used in assessing the students at entrance to a medical college. Ann King Edw Med Univ. 2012;18(3):296-9.

42. Fowell SL, Bligh JG. Recent developments in assessing medical students. Postgrad Med J. 1998;74:18-24.

43. Khan MUZ, Aljarallah BM. Evaluation of Modified Essay Questions (MEQ) and Multiple Choice Questions (MCQ) as a tool for assessing the cognitive skills of undergraduate medical students. Int J Health Sci. 2011;5(1):39-43.

44. Mol SM, Matos ASM. Uma análise sobre a taxonomia solo: aplicações na avaliação educacional. Est Aval Educ. 2019;30(75):722-47.

45. Ferreira FFG, Rocha MLPC. A Taxonomia SOLO nas teses e dissertações defendidas em programas de pós-graduação no Brasil. Revista de Matemática, Ensino e Cultura. 2020;15:32-46.