



Improvement of Faculty's Skills on the Creation of Items for Progress Testing Through Feedback to Item Writers: a Successful Experience

Melhora na Habilidade de Elaboração de Itens para o Teste de Progresso por meio de Feedback aos Autores: uma Experiência Exitosa

Pedro Tadao Hamamoto Filho¹ 

Angélica Maria Bicudo¹¹ 

KEY-WORDS

- Educational Measurement.
- Formative Feedback.
- Examination Question.
- Academic Performance.
- Faculty Medical.

ABSTRACT

The evaluation of knowledge is an important step in the teachers' work, because it is fundamental that the evaluation results be valid and reliable. Several faculty development programs have been undertaken to improve teachers' skills on item writing. Few studies, however, have explored the benefits of providing feedback to multiple-choice item writers. Progress Testing is a longitudinal assessment of students' knowledge gain. Their results are useful for generating feedback to students, teachers and medical school managers. The aim of this report is to present a successful experience of faculty development on high-quality item writing for Progress Testing through feedback about the fate of the items written by faculty members. Annually, faculties from medical schools participating in a consortium for the application of the Progress Test are asked to formulate new items to compose the test version. However, the fate of the items is not known by their authors, that is, they do not know if the items are incorporated into the test, if the items have good quality, what is the students' performance in each of the items, and what is the psychometric function of the items. In 2017, one of the schools participating in the consortium offered its item authors a feedback regarding the flaws in item writing, modification of items by the review committee and students' performance, as well as psychometric function in each item. The percentage of flawed item (items without enough quality to be selected for the test) was more than 30%. There was an inverse relation between the number of flawed item and the number of items included in the final exam, when analyzing the different content areas of the test (that is, Internal Medicine, Surgery, Pediatrics, Obstetrics and Gynecology, Public Health, and Basic Sciences). In the following year, we observed a decrease in the number of flawed item (less than 10%) and an increase in the number of eligible items to be selected for the Progress Testing. Therefore, giving feedback to item writers seems to be a good strategy for developing the faculty's ability on writing good quality items.

¹Universidade Estadual Paulista, Botucatu, São Paulo, Brazil.

¹¹Universidade Estadual de Campinas, Campinas, São Paulo, Brazil.

PALAVRAS-CHAVE

- Avaliação Educacional.
- Feedback formativo.
- Questão de prova.
- Desempenho Acadêmico.
- Docentes de Medicina.

RESUMO

A avaliação de conhecimentos é uma etapa importante do trabalho docente, pois é preciso que os resultados das avaliações sejam válidos e confiáveis. Diversos programas de desenvolvimento docente têm sido realizados para melhorar a habilidade em elaborar questões objetivas pelos professores. Poucos estudos, no entanto, exploraram os benefícios de fornecer feedback aos autores de itens de múltipla escolha. O Teste de Progresso é uma avaliação longitudinal do ganho de conhecimentos dos estudantes cujos resultados são úteis para gerar feedback aos próprios estudantes, aos professores e aos gestores das escolas médicas. O objetivo deste relato é apresentar uma experiência exitosa em gerar desenvolvimento docente na habilidade de elaborar itens de qualidade para o Teste de Progresso por meio do feedback a respeito da qualidade e do destino destes. Anualmente, os professores das escolas médicas participantes de um consórcio para aplicação do Teste de Progresso são solicitados a formular novos itens para compor a versão do teste. Porém, o destino dos itens não é conhecido por seus autores, ou seja, eles não sabem se as questões são incorporadas à prova, se elas têm boa qualidade, qual o desempenho dos estudantes em cada questão e qual o funcionamento psicométrico delas. Em 2017, uma das escolas participantes do consórcio ofereceu aos seus autores de questões uma devolutiva sobre as falhas na redação, a modificação dos itens pelo comitê revisor, o desempenho dos estudantes nas questões e o funcionamento psicométrico em cada item. A porcentagem de itens falhos, ou seja, de qualidade não satisfatória para a inclusão no teste, era superior a 30%. Houve uma relação inversa entre a quantidade de itens falhos e o número de itens incluídos na prova final, numa análise por área de conhecimento da prova (ou seja, clínica, cirurgia, pediatria, ginecologia e obstetrícia, saúde pública e ciências básicas). No ano seguinte, observamos uma diminuição no número de falhas de itens (menor que 10%) e um aumento no número de questões elegíveis a serem selecionadas para o exame de Teste de Progresso. Portanto, oferecer feedback para os redatores de questões parece ser uma boa estratégia para desenvolver a habilidade docente em elaborar itens objetivos de boa qualidade.

Received on 8/18/19

Accepted on 11/01/19

INTRODUCTION

Progress testing has been recognized as an important feedback tool for students, faculties, program directors, and schools.¹ Progress testing results help identify critical gaps in curriculum development, and point to necessary corrections.² Utilizing detailed feedback and online tools for performance monitoring, several studies have demonstrated the benefits of progress testing for student learning³⁻⁶, as well as for educational program management.⁷⁻⁹

By contrast, few studies have examined how feedback for Progress Test-item writers is generated, or even if it is offered. Previous literature has focused on item development procedures and item construction guidelines,¹⁰⁻¹² and several studies have strongly demonstrated the efficacy of faculty development programs for developing better items and more reliable exams.¹³⁻¹⁵ Furthermore, even though a large number of studies have suggested the need for faculty development programs with longitudinal monitoring of evaluative processes,¹⁶⁻¹⁸ few papers have provided feedback for item writers.^{19,20} This concern is especially important because it has been clearly demonstrated that flawed items compromise the validity and reliability of a test.²¹

The Brazilian use of progress testing has a formative characteristic. Several school consortia use progress testing to provide students with feedback; this is comprised of an exam booklet with the answer template, and appropriate commentaries, with related bibliography. This is a great opportunity for student self-assessment and, posteriorly, study guidance. However, this strategy limits the possibility of having a large item database with the repetition of previously used items. Accordingly, faculty members are often asked to write new items. And what happens to these items?

The purpose of this report is to present a school's set of initiatives that helped improve item writers' skill through feedback regarding the fate of the items written by faculty members.

EXPERIENCE REPORT**Progress Testing Methods**

Since 2005, a consortium of public medical schools, mainly from the state of São Paulo, have administered a progress testing exam to all of their students. This consortium is called NIEPAEM – *Núcleo Interinstitucional de Estudos e Práticas de Avaliação em Educação Médica* (Interinstitutional Group for Studies and Assessment Practices in Medical Education). It now includes the following schools: UNICAMP, UNESP, USP (Ribeirão Preto and Bauru campi), UNIFESP, UFSCAR, FAMEMA, FAMERP, UEL, and FURB.²²

The consortium employs a standardized blueprint, suitable for newly-qualified physicians. The exam has 120 multiple-choice questions, divided equally between six subject areas: basic sciences, internal medicine, pediatrics, surgery, obstetrics and gynecology, and public health. Every year, NIEPAEM creates a set of item orders that address the blueprint. Each school is represented at the NIEPAEM meetings by a faculty member. This faculty member is responsible for the exchange of information between their school and the others, as well as for delivering item orders to his faculty colleagues, who will be responsible for writing the required items. A single item order, therefore, might have up to ten written items. Afterwards, several specialists from the consortium schools hold a meeting to select the items that will comprise the final exam. This panel of specialists makes changes to the items to improve their quality.

For item writing, faculty members are asked to complete a standard form. The form requires the item writers to: state the item order, offer 4 answer choices (A, B, C, and D) with only one correct answer, and mark the correct answer with an "A". All items should have appropriate commentaries and related bibliographic references. Negative propositions are discouraged (i.e. "tick the wrong alternative"), as well as "all of the above are correct." Many of these practices are based on the previous Canadian and Dutch experiences.^{23,24}

Item Writing Flaws

In 2017, one of the schools from NIEPAEM had a concern regarding the quality of the delivered items. Despite always providing the item writers with guidelines for writing good items, several faculty members would send flawed items. Many items were not used; the time and efforts of item writers and specialist reviewers was, therefore, wasted. The first task, then, was to quantify the problem in terms of delivered items.

In that year, 158 item orders were delivered to the representative of each department in the school. Of these, 126 items were completed. 87 items (69%) did not have any problems, and these were taken to the specialist review board. The remaining 39 items had at least one problem. The detected problems were: five-choice questions (18), lack of bibliographic reference (10), lack of commentary (8), nonuse of "A" as the correct answer (7), non-addressing of the item theme (5), and negative proposition (1). Of the 87 items taken to the item selection meeting, 28 were accepted for use in the final exam (23.3% of the 120 items).

The Fate of the Items

After the exam had been applied, and the results were obtained, the school's local committee on progress testing submitted a report to the community. This report included an analysis of the test results, focusing on the students' performance.

The report also included individualized feedback for the item writers whose items had been used in the exam. This feedback covered: quality of writing, changes made by the specialists' panel, students' global performance on the item, psychometric indices of difficulty and discrimination, performance of the 6th-year students, and the comparison of students' performance from the same school and from different schools. Additionally, a brief explanation on the psychometric indices was added to the report. Many faculties reported great satisfaction on receiving this feedback, especially because it was the first time they had been given such information.

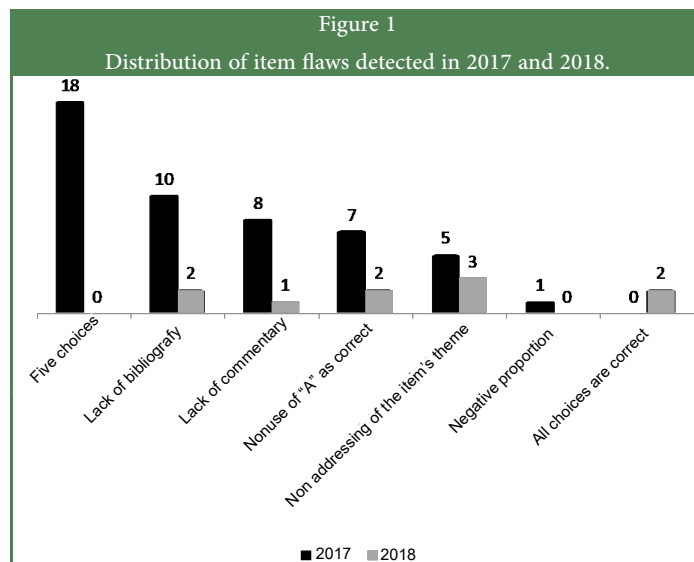
Finally, the report was delivered to all faculty members; it included information regarding flaws detected in item writing. This information also included a breakdown of the distribution of problems according to subject area: 10 on obstetrics and gynecology, 8 on public health, 3 on pediatrics and on surgery, and none on internal medicine and basic sciences. It did not take into account five-choice errors and nonuse of "A" errors, since these problems could be easily corrected by the local committee.

This analysis was a matter of concern for some departments of the school, mainly those identified with item writing problems. Meetings with the local committee were held, in order to better discuss and debate the data.

The Following Year

In 2018, the same protocol for item writing was used. Each department representative received the item orders from the local committee on

progress testing. This member requested the item writing for his colleagues. Of the 161 items ordered, 117 were delivered. Of these, 107 (94%) had no problems and were taken to the panel of specialists for the item selection meeting. Only 10 problems were detected: 3 were not sequentially ordered, 2 did not use "A" as the correct answer, 2 were missing bibliography, 2 used "all of the above are correct", and 1 did not include any commentary. Figure 1 shows the distribution of flawed item over the two years.



In 2018, of the 107 items taken to the panel of specialists for analysis, 26 were used in the final exam.

The distribution of flawed item among the six areas was: 4 on surgery, 2 on public health, 1 on internal medicine and pediatrics, and none on basic sciences and obstetrics and gynecology. Again, five choices and nonuse of "A" problems were not taken into account, since these problems could be easily corrected by the local committee.

The comparison of flawed item and accepted items showed an inverse relationship between flaws and acceptance; that is, the higher the number of detected flaws, the lower the number of items accepted for the final exam (see Figure 2).

DISCUSSION

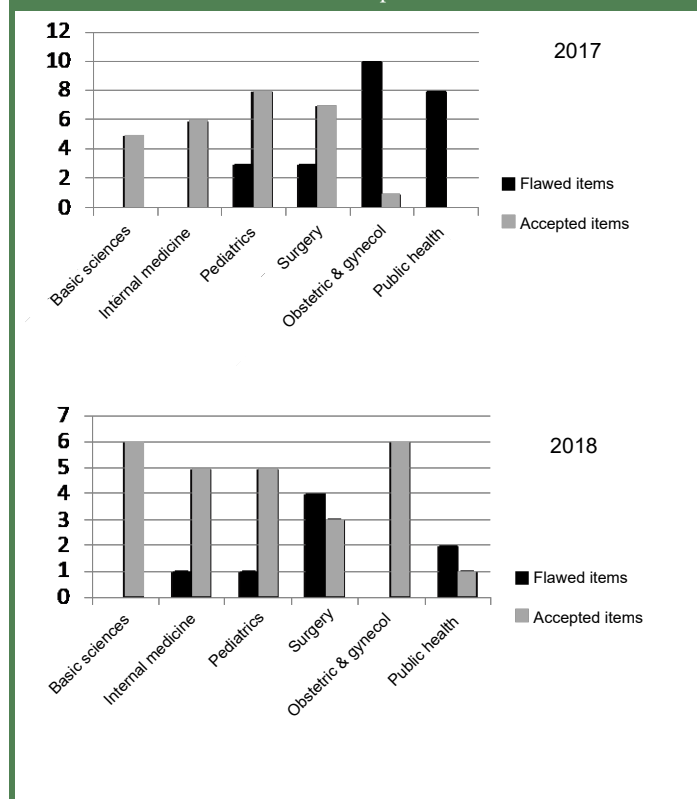
The assessment of students' knowledge is a challenge for medical education. Several domains of professionalism need to be accurately assessed and, for this purpose, multimethod and longitudinal assessments must be performed. Multiple-choice questions are helpful for the assessment of knowledge, but they are difficult to write, especially in certain content areas, such as mental health, ethics, and humanity sciences. Despite this difficulty, multiple-choice questions can provide highly reliable data, and they are routinely used all over the country.²⁵

Therefore, improving faculty members' writing skills is critical for determining a precise scale to discriminate between high and low-achieving students. Many studies have evaluated the effect of faculty development programs on item writing, but most of these comprise lower-level Kirkpatrick model,¹⁵ and there is evidence demonstrating that item writing flaws can harm the outcome of high-stake examinations.²⁶⁻²⁸

In this report, we present an institutional experience with a school-

Figure 2

Distribution of flawed and accepted items according to exam subject area. Overall, in 2018 there were fewer flawed items than in 2017. Moreover, there is an inverse relationship between the number of flawed and accepted items.



based feedback system that improved the quality of item writing for the progress testing exam. A broad evaluation, highlighting flawed item, discomfited many faculty members and departments. A certain degree of embarrassment may have motivated them to improve the quality of the requested items in the following year. This seems to have occurred especially in the areas of obstetrics and gynecology and public health. Obviously, a more direct and individualized assessment of the reactions and behavior changes of each item writer would have strengthened our observations.

In 2002, Downing recommended giving feedback to item authors²⁹; however, we could find no reports about the implementation of this suggestion, which makes our work a novel one. Accordingly, Kim et al. (2010) suggested that organic principles that reflect the item writers' trial-and-error process would benefit them more than "Dos and Don'ts Principles".³⁰

Some of our data also offer good news concerning our school's adoption of guidelines for item writing, such as the avoidance of negative propositions and "all of the above" answers, which showed a low frequency among the problems we detected, and had been identified as a particular error to be avoided.^{27,31}

Another limitation of our study is that our sample might have been biased because we enrolled only faculty members involved with progress testing item writing. We cannot overlook the fact that other faculty members might use flawed items in regular classroom examinations. However, it is important to be aware that a wave of change begins with conscientious and motivated faculties.

Finally, giving feedback to progress testing item writers will add another arm to progress testing feedback possibilities, improving its usefulness for medical school.

After this first observation, we hope that other investigators will open their eyes to the opportunities offered by the frequent requests for item writing, provide faculty members with systematic feedback regarding flaws and students' performance on each written item, and rigorously assess the impact of this set of practices to improve the quality of high-stake examinations.

CONCLUSIONS

Our work showed a significant decrease in the number of flawed items. It also showed an increase in the number of items eligible for use in progress test examinations after the disclosure of flaws and the fate of items. Giving feedback to faculty item writers seems to be a good strategy for developing faculty item writing proficiency.

REFERENCES

1. Coombes L, Ricketts C, Freeman A, Stratford J. Beyond assessment: Feedback for individuals and institutions based on the progress test. *Med Teach*. 2010; 32(6):486-490.
2. Neeley SM, Ulman CA, Sydelko BS, Borges NJ. The value of progress testing in undergraduate medical education: a systematic review of the literature. *Med Sci Educ*. 2016; 26(4):617-622.
3. Wrigley W, van der Vleuten CP, Freeman A, Muijtjens A. A systemic framework for the progress test: strengths, constraints and issues: AMEE Guide No. 71. *Med Teach*. 2012; 34(9):683-697.
4. Tio R, Schutte B, Meiboom AA, Greidanus J, Dubois EA, Bremers AJA. The progress test of medicine: the Dutch experience. *Perspect Med Educ*. 2016; 5(1):51-55.
5. Muijtjens AMM, Timmermans I, Donkers J, Peperkamp R, Medema H, Cohen-Schotanus J, Thoben A, Wernick AC, van der Vleuten CP. Flexible electronic feedback using the virtues of progress testing. *Med Teach*. 2010; 32(6):491-495.
6. Wade L, Harrison C, Hollands J, Mattick K, Ricketts C, Wass V. Student perceptions of the progress test in two settings and the implications for test deployment. *Adv Health Sci Educ*. 2012; 17(4):573-583.
7. Van der Vleuten CPM, Verwijnen GM, Wijnen WHFW. Fifteen years of experience with progress testing in a problem-based learning curriculum. *Med Teach*. 1996; 18(2):103-109.
8. Schuwirth LWT, van der Vleuten CPM. The use of progress testing. *Perspect Med Educ*. 2012; 1(1):24-30.
9. Muijtjens AMM, Schuwirth LWT, Cohen-Schotanus J, Thoben AJNM, van der Vleuten CPM. Benchmarking by cross-institutional comparison of student achievement in a progress test. *Med Educ*. 2008; 42(1):82-88.
10. Spaan M. Test and item specifications development. *Lang Assess Q*. 2006; 3(1):71-79.
11. Abdulghani HM, Ahmad F, Irshad M, Khalil MS, Al-Shaikh GK, Syed S, et al. Faculty development programs improve the quality of multiple choice questions items' writing. *Sci Rep*. 2015; 5:9556.
12. Coughlin PA, Featherstone CR. How to Write a High Quality

- Multiple Choice Question (MCQ): A Guide for Clinicians. *Eur J Vasc Endovasc Surg* 2017; 54(5):654-658.
13. Naeem N, van der Vleuten C, Alfari EA. Faculty development on item writing substantially improves item quality. *Adv Health Sci Educ Theory Pract*. 2012; 17(3):369-376.
 14. Ali SH, Ruit KG. The impact of item flaws, testing at low cognitive level, and low distractor functioning on multiple-choice question quality. *Perspect Med Educ*. 2015; 4(5):244-251.
 15. Dellenges MA, Curtis DA. Will a short training session improve multiple-choice item-writing quality by dental school faculty? A pilot stud. *J Dent Educ*. 2017;81(8):948-955.
 16. Elliot DL, Skeff KM, Stratos GA. How do you get to the improvement of teaching? A longitudinal faculty development program for medical educators. *Teach Learn Med* 1999; 11(1): 52-57.
 17. Knight AM, Cole KA, Kern DE, Barker LR, Kolodner K, Wright SM. Long-term follow-up of a longitudinal faculty development program in teaching skills. *J Gen Intern Med* 2005; 20(8):721-725.
 18. Singh T, de Grave W, Ganjiwale J, Supe A, Burdick WP, van der Vleuten C. Impact of a fellowship program for faculty development on the self-efficacy beliefs of health professions teachers: a longitudinal study. *Med Teach* 2013; 35(5):359-364.
 19. Abdulghani HM, Irshad M, Haque S, Ahmad T, Sattar K, Khalil MS. Effectiveness of longitudinal faculty development programs on MCQs items writing skills: A follow-up study. *PLoS One* 2017; 12(10): e0185895.
 20. Karthikeyan S, O'Connor E, Hu W. Barriers and facilitators to writing quality items for medical school assessments – a scoping review. *BMC Med Educ* 2019;19:123.
 21. Rush BR, Rankin DC, White BJ. The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Med Educ* 2016; 16:250.
 22. Bicudo AM, Hamamoto Filho PT, Abbade JF, Hafner MLMB, Maffei CML. Teste de Progresso em consórcio para todas as escolas médicas do Brasil. *Rev Bras Educ Méd*. 2019; 43(4):151-156.
 23. Van Der Vleuten CPM, Verwijnen GM, Wijnen WHFW. Fifteen years of experience with progress testing in a problem based learning curriculum. *Med Teach* 1996; 18(2):103-109.
 24. Blake JM, Norman GR, Keane DR, Muelle CB, Cunningham J, Didyk N. Introducing progress testing in McMaster university's problem-based medical curriculum: psychometric properties and effect on learning. *Acad Med* 1996; 71(9):1002-1007.
 25. Epstein RM. Assessment in medical education. *New Engl J Med*. 2007; 356:387-396.
 26. Haladyna TM, Downing SM. 1989. Validity of taxonomy of multiple-choice item writing rules. *Appl Meas Educ* 2:51-78.
 27. Downing SM. The effects of violating standard item writing principles on tests and students: The consequences of using flawed items on achievement examinations in medical education. *Adv Health Sci Educ* 2005; 10:133-143.
 28. Ware J, Vik T. Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations. *Med Teach* 2009; 31:238-243.
 29. Downing SM. Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference?" *Acad Med*. 2002; 77(10 Suppl):S103-S104.
 30. Kim J, Chi Y, Huensch A, Jun H, Li H, Roullion V. A Case Study on an Item Writing Process: Use of Test Specifications, Nature of Group Dynamics, and Individual Item Writers' Characteristics. *Lang Assess Q*. 2010; 7(2):160-174.
 31. Frey BB, Petersen S, Edwards LM, Pedrotti JT, Peyton V. Item-writing rules: collective wisdom. *Teaching Teach Educ*. 2005; 21:357-364.

AUTHORS' CONTRIBUTIONS

PTHF: study design, data collection and analysis, draft writing.
AMB: study supervision and final manuscript review.

CONFLICTS OF INTEREST

None to be declared.

CORRESPONDENCE ADDRESS

Pedro Tadao Hamamoto Filho
UNESP – campus de Botucatu. Distrito de Rubião Jr, s/n. Botucatu/
SP, Brasil
Departamento de Neurologia, Psicologia e Psiquiatria
pedro.hamamoto@unesp.br
(55) 14 38801220



This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.