Assessment in pediatrics clerkships: impact of strategies to solve item-sharing problems

Gabriel Faria-Costa, MD\textsuperscript{a}, Milton Severo, PhD\textsuperscript{b,∗}, Maria Amélia Ferreira, MD, PhD\textsuperscript{b}, Hercílio Guimarães, MD, PhD\textsuperscript{b}, Tiago Henriques-Coelho, MD, PhD\textsuperscript{a,c}

Abstract
Introduction: Reusing multiple-choice questions in different examinations may lead to item sharing between students. Our aim was to analyze and solve an item-sharing problem during assessment of medical students in pediatrics clerkship.

Methods: This is a 3 years prospective analysis of 5th year medical students, submitted to an examination at the end of their Pediatrics’ clerkship. In 2012, questions were reused across different clerkships. In 2013, no questions were reused in different clerkships. In 2014, no questions were reused and the review of the test was postponed to the end of the year, after all clerkships ended.

Results: In 2012, the mean score increased 1.36 points (in a scale of 0–20) per clerkship rotation, with the last clerkship having a difference of 9.5 points regarding the first one ($P < .001$). Fifty percent of this variation was due to the repetition of questions. In 2013, with a new question bank, the mean score increased 0.8 points per clerkship rotation, with a difference of 5.6 points between the last and the first rotations ($P < .024$). Finally, in 2014 there was no significant variation between clerkships.

Tests’ scores had a significant moderate correlation with students’ average course grade ($r = 0.39$, $r = 0.30$, $r = 0.48$, for 2012, 2013, and 2014, respectively). The students’ average course grade, however, did not confound the increase in tests’ scores across different clerkships.

Conclusion: The present work demonstrated an item-sharing problem among students during pediatric clerkships. An effective approach to correct this bias assessment was achieved by restricting the reuse of questions, by changing the time-point of test revision and by progressively adapting equating strategies.

Keywords: educational measurement, medical education, undergraduate

Introduction
Assessment in medical clerkships is a complex process that requires more resources than usual. Maintaining standards across groups is crucial to have an equal, reliable, valid, and fair assessment.\textsuperscript{1–3} One of the most used formats of assessment tools are Multiple Choice Questions (MCQs) tests.\textsuperscript{4} Given the difficulty of creating a new set of questions for every examination and to make tests more identical, questions are often reused, raising an old and common problem of question sharing among students.\textsuperscript{5–7} A general concern exists to maintain a safe question bank to avoid bias in the assessment process,\textsuperscript{7} but sharing among students is an inevitable problem since one can only limit the problem instead of eliminating it.\textsuperscript{8} As well, sharing questions among medical students is an ethical problem since integrity, trustworthiness, and honesty are core values for their future as health professionals.\textsuperscript{4} With the generalization of Internet use and the appearance of new electronic devices, students can easily share questions almost instantly after an examination. Facing an item-sharing problem, there must be a careful analysis to understand it and to evaluate its implications in the system that is being used.

Different strategies have been implemented in several Universities to limit this problem: (i) using a method of statistical quality control to detect possibly known items;\textsuperscript{5} (ii) seeking secure item-selection rules;\textsuperscript{10} (iii) restricting the maximum exposure rate of questions;\textsuperscript{11} (iv) overlapping and nonoverlapping rotating item banks;\textsuperscript{12–14} (v) combining previous procedures.\textsuperscript{12,13} These strategies, however, require a large sample of students and items to be applied. For small groups of students, like clerkships, a different approach is needed. Equating methods are well assigned for this scenario.\textsuperscript{16,17} Equating is a statistical method that determines the relationship between 2 or more scoring scales, to correct possible differences in tests’ difficulty.\textsuperscript{18} When equivalent tests are compared, equating can correct variations in scores that go beyond what is expected, fitting properly for medical clerkships assessment. Single group, randomly equivalent groups or nonequivalent groups with anchor test are different methods of equating.\textsuperscript{19,20}
In the present work, we prospectively analyze an item-sharing problem and the respective solving strategies, including equating, during Pediatrics clerkship assessment of 5th year medical students during 3 consecutive years. In first year there were reused questions. In second year, we used a new question bank and different questions were applied in each test but a revision time point was allowed at the end of each clerkship. In third year, we also used a new question bank but the revision time-point was postponed to end of the year, after all clerkships.

Materials and methods

Study population

In the Faculty of Medicine of the University of Porto, during 5th year, medical students are divided by 8 different pediatrics clerkships during the year. For each group of students, an assessment criteria are applied: 75% of grade is dependent on a final written test score. This organization implies that every year, 8 new tests are needed. Each test is composed by 40 MCQ with 5 different options, with only 1 correct answer. Tests are scored in a scale of 0 to 20 points, that is, the number correct answers is divided by 2.

Study design

In 2012, the examinations (MCQs tests) included reused questions. After taking the test, students only had access to the number of correct answers and we only released final scores at the end of the year. We, however, gave students the opportunity to review the test at the end of each clerkship to debate possible flawed questions. Then, we calculated the difficulty index (the percentage of students answering an item correctly) of reused questions and compared it between different number of questions’ reutilization.

In 2013, we created a new question bank and the questions were randomly used to create 8 new tests without reused questions. As occurred in 2012, after taking the test, students only knew the number of correct answers and also had the opportunity to review the test at the end of each clerkship.

In 2014, once again, we created 8 new tests without reused questions. After solving the test, at the end of each clerkship, students only knew the number of correct answers. In this year, we only allowed students to review their test at the end the academic year, after the conclusion of all 8 clerkships.

Equating approach

At the end of each year, we performed an analysis to students’ test scores across different clerkships. For this purpose we used SPSS Version 22 (Armonk, NY: IBM Corp) and R (Vienna, Austria: R Foundation for Statistical Computing). We used 1-way analysis of variance to evaluate differences in tests’ mean scores among different clerkships. We used Levene test to evaluate differences in scores’ variance among clerkships. We used a linear regression model to check if there was a trend in average course grades. We also used a new question bank but the revision time-point was postponed to end of the year, after all clerkships.

In order to check whether differences among clerkships’ tests scores could be explained by differences in the ability of students that have taken the different clerkships, we recorded students’ average course grades. Then, we adjusted the linear trend in tests’ scores across clerkships for the students’ average course grades.

Depending on this analysis, we used different equating methods in different years: we used the linear equating method\textsuperscript{18} (formula 1) when there were differences in tests’ mean scores and scores’ variance; we used the mean for equivalent groups (formula 1 assuming that \(sd1\) and \(sd2\) equals 1) method when there were only differences in tests’ mean scores.

\[
\text{Score}_1 = \frac{\text{Score}_2 - \bar{x}_2}{sd_2} \times sd_1 + \bar{x}_1
\]

where \(\text{Score}_2\) is the grade on test 2, \(\bar{x}_1\) and \(\bar{x}_2\) are the means of all students in tests 1 and 2, \(sd1\) and \(sd2\) are the standard deviations (SDs) in tests 1 and 2, and \(\text{Score}_1\) is the score the student would have had in test 1.

Results

Clerkship evaluations

In 2012, the average number of students per clerkship was 29. The analysis performed at the end of the year showed a significant increase in tests’ mean scores over the sequence of clerkships. The test’s mean score of correct answers increased 1.4 (95% confidence interval [CI]: 1.2;1.6) per clerkship (\(P\leq .001\)) (Fig. 1A). When adjusted for the students’ average course grade, this trend was still significant, 1.32 (95% CI: 1.13;1.51). The difference between the last and the first clerkship was 9.5 correct answers. The maximum mean score was 35 points and the minimum 25 points, which corresponds to a range of 11 points. The SD was also significantly different among clerkships (\(P\leq .001\)), with a mean value of 2.8 and a range of 3.2 (Table 1 and Fig. 1B). Over the 5 tests, there were 162 questions used once, 76 used twice, 40 used thrice, 24 used 4 times, 12 used 5 times, and 6 used 6 times. The difficulty index increased proportionally with the number of times that questions were used (Fig. 2). This increase in the difficulty index explained 50% of the increase in scores between subsequent clerkships (Fig. 1).

In 2013, the average number of students per clerkship was 26. With no reused questions, final analysis showed an inferior but still significant increase in tests’ mean scores over the sequence of clerkships. The test’s mean score of correct answers increased 0.8 (95% CI: 0.5;1.0) per clerkship (\(P\leq .001\)) and the difference between the last and the first clerkship was 5.6 correct answers (Fig. 1A). When adjusted for students’ average course grade, the effect was the same: a significant increase of 0.68 (95% CI: 0.43;0.93) correct answers per clerkship. The maximum mean score was 30 points and the minimum was 22 points (range of 8 points). There were, however, no significant differences in tests’ SD (\(P\leq .095\)). Although SD mean value raised to 3.7, the range decreased to 1.8 (Table 1).

In 2014, the average number of students per clerkship was 32. With no reused questions and no possibility of reviewing the test after each clerkship, there was no increase detected in tests’ mean scores. The test’s mean score of correct answers decreased −0.03 (95% CI: −0.2;0.2) per clerkship (\(P\leq .940\)) (Table 1) and the difference between the last and the first clerkship was 0.5 correct answers (Fig. 1A). When adjusted for students’ average course grade, again, no trend was detected 0.13 (95% CI: −0.04;0.30). Although not having a linear trend with the sequence of clerkships along the academic year, there was still a significant difference in tests’ mean score (Table 1). The minimum mean score was 27 points and the maximum 30 points (range of 3 points). We found no differences in tests’ SD (\(P\leq .226\)). The SD mean value was 3.3 and the range was 1.6 (Table 1).
Equating strategies

In 2012, we performed equating by the linear equating method as there were differences in both tests’ mean scores and scores’ variance. In the following years, we performed equating by the mean for equivalent groups method because there were differences only in tests’ means scores and not in scores’ variance.

Correlation between students’ scores and students’ average course grade

In all years, students’ scores had a significant moderate correlation with the average course grade ($r=0.39$, $r=0.30$, $r=0.48$, for 2012, 2013, and 2014, respectively), meaning that high-ability students had higher scores in the pediatrics test. However, as we have detailed above, after adjusting the linear trend of clerkships scores for the average course grade, the effect was always identical, showing that the course grade does not confound the linear trend.

Discussion

Assessing different groups of students at different time points and maintaining the same criteria for all assessment along the academic year is a complex labor. In the context of clerkships, this assignment becomes even more challenging, especially if the assessment tests include MCQs. Reusing MCQs, apart from making the evaluation more similar, is also useful to equate...
Non-equivalent groups.\textsuperscript{20} Reusing questions, however, always raises the concern of item sharing. In this case, both equating approach and the entire assessment procedure will be compromised. This is a controversial topic, because some works did not find differences in reused items difficulty\textsuperscript{21} or in the overall students’ performance in tests with reused items,\textsuperscript{22} whereas others point to a benefit or prior exposure to items.\textsuperscript{23} Still, few works have addressed this issue so far and even fewer have directly studied the influence of reusing questions in medical clerkships assessment. Contrarily to Herskovic,\textsuperscript{24} our results from 2012 show that reusing questions influences students’ grades, benefiting those who knew reused questions. It could be argued that these differences in students’ scores over clerkships were merely due to differences in students’ ability and not because the last clerkships knew more reused questions beforehand. To discard this hypothesis, we adjusted the linear trend of scores per clerkship for the students’ average course grade and the effect remained the same. Accordingly, the most reasonable explanation is question share among students that is reinforced by the finding that question’s reuse increased their difficulty index. Students from last clerkships scored more because they had “easier” questions and not because of higher ability. In this scenario, we could not rely equating on the compromised anchor items. Therefore, we had to adopt equating methods that do not need anchor items and since the students’ ability was not influencing scores across clerkships, we used equating methods directed to equivalent groups.\textsuperscript{20}

Tests must evaluate the acquisition of relevant knowledge with clinical significance. Consequently, different questions may be similar because they bear on the same subjects. This is highlighted in our results from 2013. Even though different questions were used, there was still a significant increase in clerkships tests’ scores that was not eliminated when adjusted for students’ average course grade. This could be explained by the access to questions after each clerkship, which could have led to spread of similar questions to other clerkships. These findings suggest that there was also an item-content sharing problem alongside the item-sharing problem itself. One of the limitations of this work is that we were not able to evaluate questions similarity. Nevertheless, we advance this explanation as plausible, and our 2014 results strengthen this explanation. In 2014, to limit item-content sharing, students only had access to the examination at the end of academic year, when all clerkships were completed. What we observed clearly contrasted with the 2 other years. The tests’ mean scores were identical between clerkships, with no tendency discerned. Clearly, the time point of the test revision plays a determinant role in students’ scores across the clerkships. Unlike other institutions, our policy allows students to review

| Table 1 | Characterization of student population |
|---------|--------------------------------------|
|         | 2012 \( N = 238 \)                  | 2013 \( N = 211 \)                  | 2014 \( N = 254 \)                  |
|         | Mean (min – max)                     | Mean (min – max)                     | Mean (min – max)                     |
| Number students by clerkship | 29 (22–37)*                          | 26 (20–33)                           | 32 (28–36)                           |
| Mean score by clerkship      | 31 (25–38)                           | 26 (22–30)                           | 28 (27–30)                           |
| SD by clerkship              | 2.8 (1.8–5.0)\textsuperscript{1}    | 3.7 (2.0–4.9)                        | 3.3 (2.6–4.2)                        |
| Cronbach alpha by clerkship  | 0.63 (0.55–0.67)                     | 0.58 (0.36–0.78)                     | 0.49 (0.16–0.72)                     |
| Linear trend of mean score per clerkship | \( \beta \) (95% CI)                | \( \beta \) (95% CI)                | \( \beta \) (95% CI)                |
| Linear trend of the mean score per clerkship adjusted for average course grade\textsuperscript{a} | 1.4 (1.2;1.6)                        | 0.8 (0.5;1.0)                        | –0.03 (–0.2;0.2)                     |

\textsuperscript{1} The mean score was significantly different between clerkships.
\textsuperscript{a} The SD was significantly different between clerkships.

Figure 2. Difficulty index variation with the number of times a question was reused.
their own tests, favoring students learning from mistakes. In the model applied in 2014 it is possible to maintain the test revision and to eliminate the effect observed in 2013.

Another interesting finding was the variation of SD across the years. Significant differences in tests SDs among clerkships were only detected in the first year. In fact, the distribution curve of SD values reveals that the maximum SD corresponds to the first clerkship and the minimum to the last. This result is consistent with the notion that the last clerkships had access to reused questions. With an easier test, the last clerkships tended to score higher and with lower variation among students. In the following years, there was no trend detected in tests’ SDs. Although SD absolute mean value increased, the SD range decreased along the years. This denotes that the tests’ difficulty was more equal between each other in the second and third years, because students tended to score with similar SD between clerkships.

Conclusion

Our results outline a strategy to control item-sharing problem among assessment along the academic year in clinical clerkships. Suppressing reused questions did not fully eliminate the trend in students’ scores. The most successful approach was the one that limited both item exposure and item-content exposure rates. Our strategy also endorsed different equating methods that best suited different scenarios. Equating prevented unfair disparities in the equality of the evaluation, especially in the first 2 years. Nevertheless, as we improve the system we also have to rethink the equating approach. It might come to the point where equating may be no longer needed. That point denotes a fair clerkship evaluating system, which we think we have achieved in the last year.

Acknowledgments
None

Conflicts of interest
None.

Previous Presentations: The results were partially presented in a short communication in the ASME Meeting 2015.

References
[1] Dalal J, Gunderman R. Standardized tests: a review. J Am Coll Radiol. 2011;8:271–274.
[2] Rom M. Grading more accurately. J Pol Sci Educ. 2011;7:208–223.

[3] McCoubrie P. Improving the fairness of multiple-choice questions: a literature review. Med Teach. 2004;26:709–712.
[4] Epstein R. Assessment in medical education. N Engl J Med. 2007;356:387–396.
[5] Gunderman R, Ladowski J. Inherent limitations of multiple-choice testing. Acad Radiol. 2013;20:1319–1321.
[6] Berlin L. The ABR “recalls” conundrum: an ethical quandary. J Am Coll Radiol. 2012;9:380–383.
[7] Lake E, Chambers D. Effects of posting examinations and answer keys on students’ study methods and test performance. J Dent Educ. 2009;73:601–613.
[8] Jiang H, Emmerton L, McKaigge V. Academic integrity and plagiarism: a review of the influences and risk situations for health students. Higher Educ Res Develop. 2013;32:369–380.
[9] Veerkamp W, Glas C. Detection of known items in adaptive testing with a statistical quality control method. J Educ Behav Stat. 2000;25:373–389.
[10] Barrada J, Olea J, Ponsoda V, et al. Incorporating randomness in the Fisher information for improving item-exposure control in CATs. Br J Math Stat Psychol. 2008;61 (pt 2):493–513.
[11] Barrada J, Abad F, Veldkamp B. Comparison of methods for controlling maximum exposure rates in computerized adaptive testing. Psychothema. 2009;21:313–320.
[12] Barrada J, Olea J, Abad F. Rotating item banks versus restriction of maximum exposure rates in computerized adaptive testing. Span J Psychol. 2008;11:618–625.
[13] Ariel A, Veldkamp B, Van der Linden W. Constructing rotating item pools for constrained adaptive testing. J Educ Measure. 2004;41:345–359.
[14] Zhang J, Chang HH, Yi Q. Comparing single-pool and multiple-pool designs regarding test security in computerized testing. Behav Res Methods. 2012;44:742–752.
[15] Huebner AR, Wang C, Quinlan K, et al. Item exposure control for multidimensional computer adaptive testing under maximum likelihood and expected a posteriori estimation. Behav Res Methods. 2016;48:1443–1453.
[16] Skaggs G. Accuracy of random groups equating with very small samples. J Educ Measure. 2005;42:309–330.
[17] Albanese AD. A general linear method for equating with small samples. J Educ Measure. 2015;52:55–69.
[18] Livingstone S. Equating Test Scores (Without IRT). Second Edition. Educational Testing Service. 2004; 11–12.
[19] Ayres S, Sunbul SO. Investigating test equating methods in small samples through various factors. Educ Sci. 2016;16:647–668.
[20] Guileria G, Gomez J. Item response theory test equating in health sciences education. Adv Health Sci Educ Theory Pract. 2008;13:1–10.
[21] Buckendahl CW, Gerrow JD. Evaluating the impact of releasing an item pool on a test’s empirical characteristics. J Dent Educ. 2016;80:1253–1260.
[22] Wagner-Menghin M, Preusche I, Schmidt I. The effects of reusing written test items: a study using the Rasch model. ISRN Educ. 2013;12:359–368.
[23] O’Neill TR, Sun L, Peabody MR, et al. The impact of repeated exposure to items. Teach Learn Med. 2015;27:404–409.
[24] Herskovic P. Reutilization of multiple-choice questions. Med Teach. 1999;21:430–431.
[25] Kreter G, Green J, Lenoch S. The overall impact of testing on medical student learning: quantitative estimation of consequential validity. Adv Health Sci Educ Theory Pract. 2013;18:835–844.