MODESTUM
OPEN ACCESS

EURASIA Journal of Mathematics, Science and Technology Education, 2018, 14(7), 3301-3309
ISSN:1305-8223 (online) 1305-8215 (print)
Research Paper https://doi.org/10.29333/ejmste/91681

Does Repetition of the Same Test Questions in Consecutive Years Affect their Psychometric Indicators? – Five-year Analysis of In-house Exams at Medical University of Warsaw

Mariusz Panczyk [1*], Aleksander Zarzeka [1], Marcin Malczyk [2], Joanna Gotlib [1]

[1] Division of Teaching and Outcomes of Education, Faculty of Health Science, Medical University of Warsaw, Warsaw, POLAND
[2] University Exams Office of Medical University of Warsaw, Warsaw, POLAND

Received 10 January 2018 ▪ Revised 30 April 2018 ▪ Accepted 4 May 2018

ABSTRACT

Aim of study: Evaluation of the re-used test questions on the impact of psychometric indicators of test items in examinations in cardiology in Physiotherapy at the Medical University of Warsaw (MUW).

Materials and Methods: A case study based on analysis of 132 five-option (528 distractors) multiple-choice questions (MCQs) developed at MUW included in five in-house exams. Questions repeated at least twice during the period considered and constituted 42.4% of all MCQs. Each MCQ was assessed on the basis of the following three indicators: difficulty index (DI), discrimination power (DP), and the number of non-functioning distractors (N-FD). The change in psychometric indicators of test items was assessed using Krippendorff alpha coefficient (α_k).

Results: Together with each MCQs repetition, a decrease in the number of questions that would maintaining the analogical DI value towards the initial level of easiness was observed. However, the level of DI compliance was significantly higher, even when there were five consecutive repetitions (coefficient α_k for the consecutive repetitions was 0.90, 0.85, 0.78 and 0.75). N-FD number in consecutive repetitions remained on a satisfactory level (good and very good compliance), although there was a significant decrease in this range when there were three or more repetitions (coefficient α_k was 0.80, 0.69, 0.66 and 0.65, respectively). Whereas the level of similarity as for DP for consecutive repetitions was significantly lower in comparison with those noted for DI and DE (DP coefficient α_k was 0.28, 0.23, 0.25 and 0.10, respectively).

Conclusions: The observed change in the initial values of psychometric indicators together with consecutive use of the same MCQs confirms the examiners' concerns as for the progressive wear of the bank of test questions. However, the level of psychometric MCQs values loss, especially in the area of the easiness and the number of non-functioning distractors was not drastic. It appears that the level of MCQs spread among students of consecutive years is not too high, at least within two consecutive years.

Keywords: educational measurement, test questions, differentiation power, ease, health sciences, students

BACKGROUND

"May I use the same test question during another exam round"? Every teacher who applies tests in their teaching as a method that checks students' knowledge faces this question while evaluating their students' educational

✉ mariusz.panczyk@wum.edu.pl **(*Correspondence)** ✉ aleksander.zarzeka@wum.edu.pl
✉ marcin.malczyk@wum.edu.pl ✉ joanna.gotlib@wum.edu.pl

Contribution of this paper to the literature

- Multiple use of the same MCQs results in reduced differentiation power of examination tests prepared using the same pool of questions.
- By limiting the use of the same pool of MCQs more than twice in a row it is possible to maintain an adequate number of functioning distracters, the difficulty and the differentiation power of MCQs from the bank of test questions.
- The rotational and rational use of the limited resources of MCQs while preparing and administering an in-house exam may extend the "life" of the bank of test questions.

achievements. Long-term analyses of test question pools offer a closer look at how psychometric indicators of test questions, particularly those used more than once in subsequent rounds of in-house exam, change over the years.

For an average teacher it is difficult to create test questions (e.g. multiple-choice questions (MCQs) with very good psychometric indicators. The necessity to write new MCQs each year poses an additional difficulty. The question as to whether to recycle questions is an issue any organization administering tests or examinations must address (Kim, 2013). Once MCQs are re-used, the probability of their content becoming publicly available is higher, which gives advantage too those who have access to it (And in the age of the Internet, once they do leak out, there's a good chance that someone will post the information online (Somin, 2011)). Additionally, there is also concern about question repetitions in situations when a student failed their exam and is obliged to retake it.

Therefore, in many cases re-using MCQs is required on the grounds of the necessity to provide comparable conditions of evaluation (equivalence in consecutive rounds in the scope of fairness and validity of measurement) (Yang, Lee, & Park, 2018). In the literature, data on the subject of the influence of sharing banks of test questions can be found. It has been demonstrated that public access to such banks does not have to negatively influence fairness and validity of an exam (Wagner-Menghin, Preusche, & Schmidts, 2013; Wood, 2009; Wood, St-Onge, Boulais, Blackmore, & Maguire, 2010; Yang et al., 2018). However, despite these reports, there is currently no consensus in the field of unambiguous evaluation of the influence of sharing content of questions on the effectiveness of examination procedure (Yang et al., 2018).

Assessing students using tests including questions that have already been used, bears a risk that responses obtained in this way will not comply with the criteria of good educational measurement (T. J. Wood, 2009). It may be assumed that the more often a question is asked, the more plausible it will be that this question will become a part of the so-called "public domain" and will as such bring additional benefits to students who had access to such questions. Additional problem associated with re-using the MCQs it is the occurrence of practice effects in the group of exam retaking students who are exposed to the questions which they had already known from the first exam. (O'Neill, Sun, Peabody, & Royal, 2015). These factors may contribute to an increase in the ease and reduced varying ability of the test. In the literature the described above phenomenon of changes in the properties of individual test items after repeated exposure is defined as Item Parameter Drift (IPD) (Krause, 2012). The main parameter applied in the detection of IPD is the increase in the easiness of a question. The data concerning the easiness of the level of repeated questions (retest) published so far show that even though a re-used item is characterized by a greater easiness (Cates, 1982; Hertz & Chinn, 2003; O'Neill, Lunz, & Thiede, 2000; Raymond, Neustel, & Anderson, 2009; Wood, 2009), this change is slight and remains within the limits of error measurement for a given test (O'Neill et al., 2000). Moreover, the results published by O'Neill et al. (2015) on the influence of the re-used set of test questions during the American Board of Family Medicine's certification examination show that even if in some cases using test questions again may bring benefits to the test takers, the generally observed change is by and large connected with guessing than with the prior knowledge of questions.

Due to the lack of possibility of constant supply of new MCQs to the bank of test questions, it seems rational to include control of psychometric properties of MCQs in the procedure of ensuring the quality of evaluation. The quoted findings allow to assume that re-using the same test question in another evaluation round should not significantly worsen the psychometric properties of MCQs. Yet it seems appropriate to ask after how many times of using a given MCQ, its psychometric properties worsen so much that it fails to fulfil the assumed criteria. Answering this question requires compiling and assessing psychometric properties applied for MCQs and that should not be solely limited to the difficulty index (DI). Evaluation of the bank of test questions should include also assessment of the discrimination power (DP) and the number of non-functioning distractors (N-FD). Following the changes that take place in psychometric properties of MCQs in consecutive sessions or test rounds, should give an opportunity of inclusion of defence mechanisms that would protect against the excessive use of questions and lowering the quality of assessment. Additional advantage of carrying out such control could be applying the results of test questions evaluation in the improvement of faulty or "overused" MCQs (Considine, Botti, & Thomas, 2005).

Based on available premises in the literature, three hypotheses, verified in the course of the study, have been formulated:

1st hypothesis – Re-using MCQs in subsequent years does not influence DI and N-FD indicators.

2nd hypothesis – Re-using MCQs in subsequent years negatively influences DP indicators.

3rd hypothesis – Indicators DI, N-FD and DP are at a similar level regardless of single or multiple use of MCQs.

AIM OF STUDY

The present study aimed to assess the influence of re-used test questions on the change in psychometric indicators of test items (DI, N-FD, and DP) in examinations in Cardiology in Physiotherapy at the Medical University of Warsaw (MUW).

MATERIAL AND METHODS

Context

The test was carried out at MUW that is one of the biggest medical universities in Poland. Among other faculties, MUW educates students at the Faculty of Physiotherapy, where studies are conducted on the BA as well as MA levels. Within the two-year programme of MA studies, one of nine subjects of specialized education during the first year is functional diagnostics and scheduling rehabilitation in cardiology. Learning outcomes in this subject include issues connected with the structure, functioning and pathological changes in blood circulation and structural changed caused by the illness. This knowledge is then applied in functional diagnostics and in physiotherapeutic treatment. As part of the subject, students have seminars in groups of up to 20 people (10h), and also clinical exercises in groups of 6 people (30h). Additionally, students are obliged to undergo internship in the field of clinical physiotherapy in cardiology (20h).

The level of assumed learning outcomes for students is evaluated using a test comprising 50 MCQs elaborated and developed by the lecturers from Cardiology Clinic, Department of Physiotherapy, 2nd Faculty of Medicine MUW, which is a unit responsible for teaching this subject. MCQs established for the purpose of the exam are evaluated by the head of the clinic who makes the final approval of the task pool included in the bank of test questions. The bank of tasks is administered by Examinations Office which also organizes all the exams at the MUW. The course and the conditions of student assessment are described in the procedure and are regulated by the Rector of the MUW (Decree No. 93/2014).

Data Collection

Pool of test questions prepared for exams in the field of knowledge in cardiology between the academic years of 2008/09 and 20212/13 included 132 five-option MCQs (516 distractors), 76 of which were MCQs (57.6%) that were used once only, and 56 MCQs (42.4%) used at least twice. The percentage of MCQs not used in testing before was changeable and was 100% (2008/09), 62% (2009/10), 28% (2010/11), 32% (2011/12), and 36% (2012/13).

In the tested period, 100 new physiotherapy students took the cardiology exam each year (498 students in total throughout a five-year period). Each student could take the test once only. The exam was carried out in a traditional, pen and paper form. The conditions during the exam were comparable as far as time is concerned (60 minutes), the number of test versions (2 versions different in the order of questions in a set) and the number of test takers. Calculating score was done automatically, using a test card reader (scanner) and computer software TESTY version 7 ("Testy komputerowe", Copyright © 1994-2014 by Sławomir Zalewski, licence issued for MUW).

Psychometric Indicators

While evaluating psychometric properties of MCQs, a concept based on the measurement evaluation included in Classical Test Theory (also known as classical true score theory) was applied. CTT is a simple linear psychometric model describing how measurement errors may influence the observed result (Schuwirth & van der Vleuten, 2011). Traditionally, CTT uses two indicators in evaluating psychometric measurement in form of a test: difficulty index (DI) and discrimination power (DP) (Erguven, 2013). Additionally, for every MCQ from the pool of questions, there was also a number of non-functioning distractors (N-FD) (Tarrant, Ware, & Mohammed, 2009).

Statistical Analysis

In evaluating changes in the values of individual psychometric indicators for 53 MCQs which were used during a five-year period at least twice, Krippendorff alpha coefficient (α_k) was used (Krippendorff, 2012). Coefficient α_k developed to measure the agreement among observers, coders, judges, raters, or measuring instruments. α_k emerged in content analysis but is widely applicable wherever two or more methods of generating data are applied

Table 1. Compliance of difficulty index value for the correct option and the frequency of selection of the distractors for questions with different number of repetitions

Questions with a number of repetitions	Krippendorff alpha coefficient		
	Difficulty index	Discrimination power	Number of non-functioning distractors
Twice	0.90	0.28	0.80
Three times	0.85	0.23	0.69
Four times	0.78	0.25	0.66
Fivefold	0.75	0.10	0.65

Table 2. Change of psychometric indicators' value for MCQs repeated twice or three times

Psychometric indicators	1st vs 2nd			1st vs 3rd		
	Increase	No change	Reduction	Increase	No change	Reduction
Difficulty index	32 (57%)	6 (11%)	18 (32%)	23 (68%)	2 (6%)	9 (27%)
Discrimination power	27 (48%)	3 (5%)	26 (47%)	15 (44%)	0 (0%)	19 (56%)
Number of non-functioning distractors	8 (14%)	40 (72%)	8 (14%)	8 (23%)	20 (59%)	6 (18%)

to the same set of objects, units of analysis, or items. This coefficient is the most universal compliance coefficient as it has no limitations concerning: type of a measurement scale, value number / category within scale, number of repeated measurements and the minimum number of evaluated cases. Its undoubted advantage lies in its resistance to the lack of data. α_k coefficient assumes values between -1.00 and +1.00, whereas 0.00 means compliance with the level of cases and +1.00 means perfect compliance (Krippendorff, 2012).

Additionally, the number of MCQs was assessed for which psychometric indicators were altered after two and three repetitions. Values of psychometric indicators for 76 MCQs were established, if MCQs were used once only, and these values were compared with the values of questions used ≥ 2 times. Due to the different number of people in groups and no regular spread accessible for comparison of the pool of questions used once and twice, a non-parametric Mann Whitney U test was used and the effect size was evaluated by calculating biserial correlation coefficient (r_b). The conditions of using Mann Whitney U test were checked by assessing the similarity of dispersion of a dependent variable in both compared groups (Ansari-Bradley dispersion test) (Nachar, 2008).

STATISTICA software, version 12.5, was used in calculations, together with an additional module "Zestaw PLUS" (StatSoft, Inc.) in compliance with the licence issued for MUW. For each analysis, the level of statistical significance assumed *a priori* was $\alpha = 0.05$.

RESULTS

Verification of the 1st Hypothesis

The analysis of value similarity indicators for the repeated questions shown that there is a very good or good compliance in this area for DI and N-FD indicators (**Table 1**). In case of DI indicators, with every repetition of MCQs, a decrease in the number of questions maintaining the analogical DI value was observed in comparison with the initial level of easiness. However, the level of DI compliance was high enough even with five consecutive repetitions ($\alpha_k > 0.70$). The N-FD number in consecutive repetitions remained on a satisfactory level (good or very good compliance), although there was an observed significant drop in this area for three or more repetitions (**Table 1**).

While evaluating the influence of a re-used MCQ on psychometric indicators' value, the change in the easiness was checked as well as the number of distractors between the first and the second use (1st vs 2nd), and between the first and the third use (1st vs 3rd) (**Table 2**). An increase in the DI value was observed for a significant percentage of the repeated MCQs both while comparing the 1st vs 2nd and 1st vs 3rd (DI increase for 57% MCQs for two repetitions and 68% MCQs for three repetitions). Whereas in case of changing the number of ineffective distractors while comparing 1st vs 2nd , 72% MCQs remained without a change, and for 1st vs 3rd 59% MCQs.

Verification of the 2nd Hypothesis

Findings concerning DP indicator remain in contrast to these as there was no or very little compliance. For consecutive repetitions, coefficient α_k was significantly lower in comparison with those noted for DI and DE. Also, a low value of α_k noted for questions with five repetitions was characteristic (**Table 1**). Additionally, a lower differentiating ability was noted for MCQs while comparing 1st vs 2nd and 1st vs 3rd (lowering the DP value for 47% MCQs for two repetitions and for 56% MCQs for three repetitions) (**Table 2**).

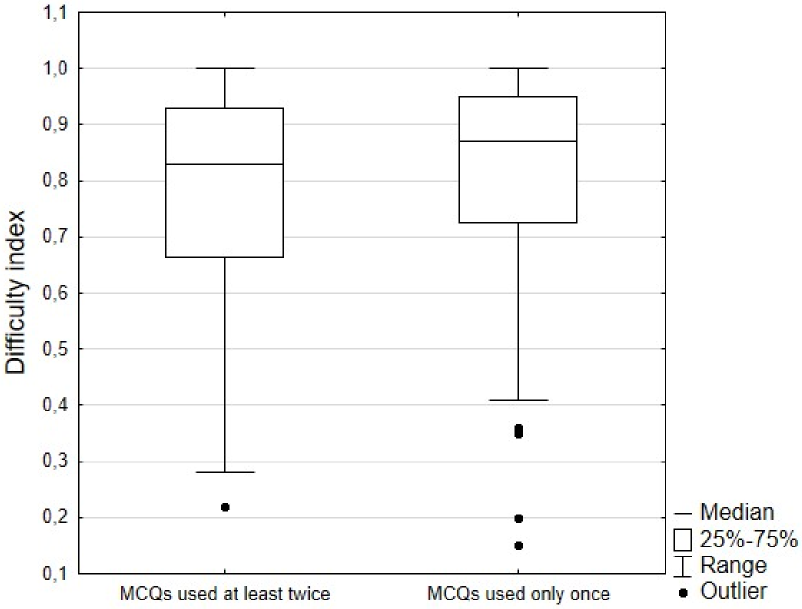

Figure 1. Comparison of the difficulty index value calculated after the first use of the MCQs pool and used at least twice as opposed to those used once only

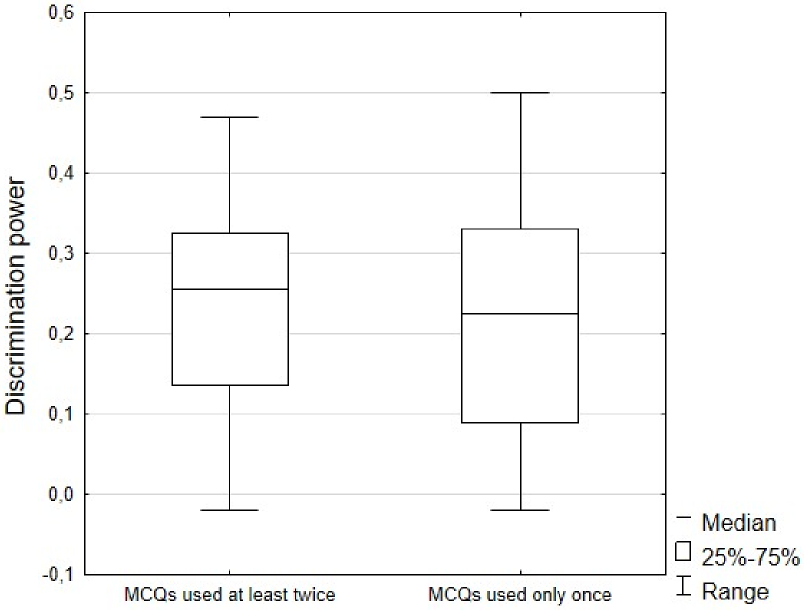

Figure 2. Comparison of the discrimination power value calculated after the first use of MCQs pool and used at least twice as opposed to those used once only

Verification of the 3rd Hypothesis

While comparing the level of easiness of MCQs for questions repeated at least twice and those used once only, there was no statistically significant differences noted between these two pools of MCQs (Mann Whitney test, U=1723.0, *P*=0.365, r_b=0.09; **Figure 1**). Similarly, there were no differences while comparing discrimination power (Mann Whitney test, U=987.5, *P*=0.558, r_b=0.07; **Figure 2**). Additionally, no statistically significant differences were noted while comparing the number of non-functioning distractors (Mann Whitney test, U=819.5, *P*=0.849, r_b=0.02; **Table 3**).

Table 3. Comparing the number of non-functioning distractors calculated after the first use of an MCQs pool that were used at least twice as opposed to those used once only

Number of non-functioning distractors	MCQs used at least twice		MCQs used only once	
	N	%	N	%
Four	15	26.8%	11	32.4%
Three	25	44.6%	11	32.4%
Two	10	17.9%	4	11.7%
One	5	8.9%	4	11.7%
Zero	1	1.8%	4	11.7%

DISCUSSION

Despite the large number of studies and guidelines for the proper creation and development of MCQs (Case & Swanson, 2002) our current knowledge of "aging" of the test questions which were used several times for evaluation of the same or different student groups is insufficient (O'Neill et al., 2015). Although there are research results on the influence of public availability of large banks of test questions (Gilmer, 1989; Park & Yang, 2015; Wagner-Menghin et al., 2013; Wood, 2009; Yang et al., 2018) on exam results, the conclusions from this type of research cannot be directly transferred to in-house exams, which are conducted on a much smaller scale than large country-wide standardized exams. Among teachers, there is a conviction that the use of the same MCQs for next year exam will contribute to an increase in its ease which will result in inflated scores.

Once the text of the MCQs is transferred from one student to another, it may become the cause of dishonesty, as some students have access to the items, while others do not. What teachers conducting in-house exam may do in order to lessen the risk connected with sharing the content of the questions? One possible solution is sharing the MCQs used so far with all students, so that all of them have equal chances (Somin, 2011). This strategy undoubtedly allows maintaining fairness of the exam, but at the same time increases workload, as each year a new set of MCQs needs to be prepared.

The obtained results confirmed authors' assumptions as to the change in value of psychometric indicators of MCQs after using them at least twice in different groups of students. Fulfilling the assumption that MCQs used ≥2 initially had similar values of psychometric indicators to the questions used once demonstrated that reused MCQ influenced on the values of the indicators in a manner which manifests in a decrease of compliance of obtained values of psychometric indicators. While in the case of the double use of an MCQ pool, the ease and the number of non-functional distractors retains a high degree of similarity to the first use, the next rounds in which the MCQs are exposed significantly influence the compliance of obtained indicator DI and N-FD values. The significant deterioration of compliance is also observed in case of indicator DP but in this case after the second use of an MCQ pool the compliance of registered values was significantly decreased for this indicator ($\alpha_k < 0.30$).

A more thorough assessment of changes in the value of three psychometric indicators for repeated MCQs allowed to determine the direction of changes in subsequent exposures of items in a new group of students. In all cases the double usage of MCQs had less negative impact on DI, DP and N-FD than three times usage of the same MCQs. The above findings show the existing relationship between the number of MCQ exposures and noted value of psychometric indicators.

The observations described above are related to the reuse of MCQs in different groups of students. Therefore, there is an opportunity to exchange information regarding examination questions between those groups. However, it is not known to which extent the used MCQs are remembered so well, they can be effectively passed to students taking examinations in consecutive rounds. The available literature provides us with some knowledge about the effects of subsequent exposure of the same candidate to tasks previously used in the evaluation. These findings refer to the situation when the content of the tasks on the first try was remembered and used by the same person during the retaking of the examination. It is therefore a scheme of repeated measurement on the same group of candidates (retest). Boulet, McKinley, Whelan, and Hambleton (2003) noticed that candidates retaking the exam using a standardized patient (clinical skills assessment) did not obtain significantly better or worse results in comparison to the first examination. Similar results were obtained by Raymond et al. (2009). On the other hand Wood (2009) noticed that students retaking an exam obtained higher scores but this findings applied to both the pool of new and reused questions. O'Neill et al. (2015) observed that the results obtained by the students retaking the examination were slightly better than received at the first approach. However, the estimated effect size for the difference in scores obtained for unique questions and reused questions was small (O'Neill et al., 2015). Also Swygert, Balog, and Jobe (2010) registered some growth in scoring obtained in retaken examination for the United States Medical Licensing Examination series Step 2 Clinical Skills. Greater number of received points did not however depend on whether or not the student knew the question (Swygert et al., 2010).

The above data, even though it represents measurements of the retest type, gives reason to believe that the repeated use of previously disclosed questions does not have to affect the increased results of the next evaluation. If the observed increase in results of retaken exams is similar in terms of new and reused tasks, it can be argued that the degree of the memorising of the questions is not as significant as it might appear. The differences observed at the MUW regarding the ability of differentiation and growth ease of MCQs between two and three times usage show that solutions protecting from the excessive use of the question bank should be introduced.

Certainly an important factor that could reduce the risk of excessive "aging" of questions is the use of tasks with high differentiation parameters and high efficiency. It is usually difficult to remember such questions in detail as far as the content of header (stem) and particular answer options are concerned. In addition, the introduction of new types of questions (such as Extended Matching Questions or Short Answer Questions) reduces the chance that the questions memorized by students will be a valuable source of information for subsequent student years. (Mujeeb, Pardeshi, & Ghongane, 2010; Oyebola et al., 2000).

Open banks of exam questions are published in numerous countries and every student can become familiar with their contents and thus prepare for the final exam or mid-term test (Considine et al., 2005; Hansen & Dexter, 1997; Park & Yang, 2015; Yang et al., 2018). It needs to be noted, however, that open task banks usually contain a few thousand of items (e.g. The Dutch Progress Test is in the public domain, but it contains over 10.000 items grouped in 19 disciplines and 17 categories (Tio et al., 2016)). Academic textbooks also use test questions as a teaching tool in order not to check knowledge but to teach (Masters et al., 2001). Thus, it seems that the presence of the "student market" should not discourage examiners from re-using MCQs. However, the conditions that in this case must be met are: (1) constant control of psychometric indicators of MCQs, (2) rotary and rational usage of available resources of the tasks bank and (3) creation of new types of questions which could be used in computer examination. Furthermore, the adding to the procedures of university assessment quality management the following principles: constant control and analysis of trends in changes of MCQ psychometric properties; the obligation of taking appropriate actions by test questions developers which should be aimed at improving operational efficiency of distractors and elimination of defective construction of MCQs. Some options include creating new test questions ever year or allowing a sufficient amount of time (2-3 years) between question re-use. Although there are several suggested solutions, the question as to whether test items should be reused and recycled remains an unanswered one (Kim, 2013).

LIMITATIONS OF STUDY

The main limitation of the presented case study was analysis of psychometric indicators of reused MCQs using the concept of evaluation of measurement described CTT. For more accurate description of the complex characteristics of the measurement characteristics of the hypothetical feature dominating the responses given in the test, the more appropriate would be the usage of a model based on Item Response Theory (IRT). However, at the item level, the CTT model is relatively simple. CTT does not invoke a complex theoretical model to relate an examinee's ability to success on a particular item. Instead, CTT collectively considers a pool of examinees and empirically examines their success rate on an item (Erguven, 2013). An important argument for the use in the presented work of psychometric indicators resulting from CTT is a small number of items. For the correct item parameters estimation suffice the number of items in the range of 200-500, whereas in the case of IRT generally required number is >500 items (Erguven, 2013; Hambleton & Jones, 1993).

CONCLUSIONS

The change of the output values of psychometric indicators together with the reuse of the same MCQs indicate a gradual wear of bank of test questions. On the one hand, after the double use, the degree of loss of the desired properties of psychometric MCQs within the ease and the number of non-functional distractors was small, contrasting with a clear reduction in power of differentiation. On the other hand, the three times used MCQs were characterized by already considerable deterioration of psychometric properties. It appears that the prevalence of MCQs contents among students of two subsequent student years is not large enough not to be able to use twice some of a pool of the same MCQs exams assessing the knowledge. It is necessary to continue studies with the aim to determine the impact of the reuse of the same MCQs however not year after year but every two or three years. Moreover, it is necessary to extend of the research and using a large number of items what would allow the use of non-linear modelling based on the IRT.

REFERENCES

Boulet, J. R., McKinley, D. W., Whelan, G. P., & Hambleton, R. K. (2003). The effect of task exposure on repeat candidate scores in a high-stakes standardized patient assessment. *Teaching and Learning in Medicine, 15*(4), 227-232. https://doi.org/10.1207/S15328015TLM1504_02

Case, S. M., & Swanson, D. B. (2002). *Constructing Written Test Questions For the Basic and Clinical Sciences* (3rd ed.). National Board of Medical Examiners.

Cates, W. M. (1982). The efficacy of retesting in relation to improved test performance of college undergraduates. *The Journal of Educational Research, 75*(4), 230-236. https://doi.org/10.1080/00220671.1982.10885386

Considine, J., Botti, M., & Thomas, S. (2005). Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian, 12*(1), 19-24. https://doi.org/10.1016/S1322-7696(08)60478-3

Erguven, M. (2013). Two approaches to psychometric process: Classical test theory and item response theory. *Journal of Education, 2*(2), 23-30.

Gilmer, J. S. (1989). The effects of test disclosure on equated scores and pass rates. *Applied psychological measurement, 13*(3), 245-255. https://doi.org/10.1177/014662168901300303

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 253-262. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hansen, J. D., & Dexter, L. (1997). Quality multiple-choice test questions: Item-writing guidelines and an analysis of auditing testbanks. *Journal of Education for Business, 73*(2), 94-97. https://doi.org/10.1080/08832329709601623

Hertz, N., & Chinn, R. (2003). *Effects of question exposure for conventional examinations in a continuous testing environment.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago.

Kim, H. (2013). Reuse, Reduce, Recycle... Test Questions? Retrieved from http://edtheory.blogspot.kr/2013/11/reuse-reduce-recycletest-questions.html

Krause, J. (2012). *Assessment of item parameter drift of known items in a university placement exam.* Arizona State University.

Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*. Newbury Park: Sage.

Masters, J. C., Hulsmeyer, B. S., Pike, M. E., Leichty, K., Miller, M. T., & Verst, A. L. (2001). Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. *Journal of Nursing Education, 40*(1), 25-32. https://doi.org/10.3928/0148-4834-20010101-07

Mujeeb, A., Pardeshi, M., & Ghongane, B. (2010). Comparative assessment of multiple choice questions versus short essay questions in pharmacology examinations. *Indian Journal of Medical Sciences, 64*(3), 118-124. https://doi.org/10.4103/0019-5359.95934

Nachar, N. (2008). The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. *Tutorials in Quantitative Methods for Psychology, 4*(1), 13-20. https://doi.org/10.20982/tqmp.04.1.p013

O'Neill, T. R., Sun, L., Peabody, M. R., & Royal, K. D. (2015). The Impact of Repeated Exposure to Items. *Teaching and Learning in Medicine, 27*(4), 404-409. https://doi.org/10.1080/10401334.2015.1077131

O'Neill, T., Lunz, M. E., & Thiede, K. (2000). The impact of receiving the same items on consecutive computer adaptive test administrations. *Journal of Applied Measurement, 1*(2), 131-151.

Oyebola, D. D., Adewoye, O. E., Iyaniwura, J. O., Alada, A. R., Fasanmade, A. A., & Raji, Y. (2000). A comparative study of students' performance in preclinical physiology assessed by multiple choice and short essay questions. *African Journal of Medicine and Medical Sciences, 29*(3-4), 201-205.

Park, Y. S., & Yang, E. B. (2015). Three controversies over item disclosure in medical licensure examinations. *Medical Education Online, 20*(1), 28821. https://doi.org/10.3402/meo.v20.28821

Raymond, M. R., Neustel, S., & Anderson, D. (2009). Same-Form Retest Effects on Credentialing Examinations. *Educational Measurement: Issues and Practice, 28*(2), 19-27. https://doi.org/10.1111/j.1745-3992.2009.00144.x

Schuwirth, L. W., & van der Vleuten, C. P. (2011). General overview of the theories used in assessment: AMEE Guide No. 57. *Medical Teacher, 33*(10), 783-797. https://doi.org/10.3109/0142159X.2011.611022

Somin, I. (2011). The Perils of Reusing Questions from Past Exams. Retrieved from http://volokh.com/2011/01/18/the-perils-of-reusing-questions-from-past-exams/

Swygert, K. A., Balog, K. P., & Jobe, A. (2010). The impact of repeat information on examinee performance for a large-scale standardized-patient examination. *Academic Medicine, 85*(9), 1506-1510. https://doi.org/10.1097/ACM.0b013e3181eadb25

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Medical Education, 9*, 40. https://doi.org/10.1186/1472-6920-9-40

Tio, R. A., Schutte, B., Meiboom, A. A., Greidanus, J., Dubois, E. A., & Bremers, A. J. (2016). The progress test of medicine: the Dutch experience. *Perspectives on Medical Education, 5*(1), 51-55. https://doi.org/10.1007/s40037-015-0237-1

Wagner-Menghin, M., Preusche, I., & Schmidts, M. (2013). The effects of reusing written test items: A study using the Rasch model. *ISRN Education, 2013*. https://doi.org/10.1155/2013/585420

Wood, T. J. (2009). The effect of reused questions on repeat examinees. *Advances in Health Sciences Education, 14*(4), 465-473. https://doi.org/10.1007/s10459-008-9129-z

Wood, T. J., St-Onge, C., Boulais, A.-P., Blackmore, D. E., & Maguire, T. O. (2010). Identifying the unauthorized use of examination material. *Evaluation & the Health Professions, 33*(1), 96-108. https://doi.org/10.1177/0163278709356192

Yang, E. B., Lee, M. A., & Park, Y. S. (2018). Effects of test item disclosure on medical licensing examination. *Advances in Health Sciences Education, 23*(2), 265-274. https://doi.org/10.1007/s10459-017-9788-8

http://www.ejmste.com