Higher Education Studies; Vol. 10, No. 2; 2020
ISSN 1925-4741 E-ISSN 1925-475X
Published by Canadian Center of Science and Education

Should Items and Answer Keys of Small-Scale Exams Be Published?

Hüseyin Selvi[1]

[1]Medical Faculty, Mersin University, Mersin, Turkey

Correspondence: Hüseyin Selvi, Medical Faculty, Mersin University, Mersin, Turkey, Çiftlikköy Campus, Mezitli, Mersin, Turkey. Tel: 90-324-361-0001. E-mail: hsyn_selvi@yahoo.com.tr

Orcid id: https://orcid.org/0000-0002-3513-0003

Received: February 15, 2020 Accepted: March 14, 2020 Online Published: March 24, 2020

doi:10.5539/hes.v10n2p107 URL: https://doi.org/10.5539/hes.v10n2p107

Abstract

This study aimed to examine the effect of using items from previous exams on students' pass-fail rates and on the psychometric properties of the tests and items.

The study included data from 115 tests and 11,500 items used in the midterm and final exams of 3,910 students in the preclinical term at the Faculty of Medicine from 2014 to 2019. Data were analyzed using descriptive statistics related to the total test scores, item difficulty and item discrimination values, and internal consistency values for reliability. The Shapiro-Wilks test was used to evaluate the distribution structure, and t test were used to analyze the differences between groups.

The findings showed that the mean item repetition rate from 2014 to 2019 ranged from 16.98% to 39.00%. The total score variance decreased significantly as the percentage of test items increased. There was a significant, moderately positive relationship between the percentage of repeated test items and the number of students eligible to pass their grades. Item difficulty values obtained from initial item use were significantly lower than those obtained from repeated item use.

We conclude that test items and answer keys should not be published by test makers unless they have the means such as the infrastructure, budget, and personnel to develop new items in place of the ones previously published in test banks.

Keywords: rote learning, item disclosure, small-scale exams, reused item, item banking

1. Introduction

Cognitive skills that are tried to be gained to individuals by education and should be measured accordingly are listed by Krathwohl (2002) as the levels of remembering, understanding, applying, analyzing, evaluating and creating from the lowest level to the highest level. This ranking shows that the simplest skill group is found at the level of remembering. It is not possible to acquire the highest-level skills without gaining the lower-level skills. In other words, it is not possible to attain the levels of comprehension, application, analysis, evaluation, and synthesis without achieving the level of remembering. Therefore, the acquisition of knowledge at the level of remembering is a significant step in education. However, simply recalling information is not a goal in itself but a means for achieving higher-level skills. Hence, the tests used to measure these skills should not only focus on measuring the level of remembering. As doing so could push students to merely memorize information only. This may cause the skills that are tried to be gained to individuals by education not to go beyond the level of remembering (Erkuş, 2006).

In addition, the items (focusing on concepts, events, or situations) that will be used in measuring high-level cognitive skills should be new to the respondent, since the test cannot go beyond measuring recall if the individuals have encountered these items before.

The fact that students encounter with the items before the exam often creates problems in educational programs. The curriculums relevant to the skills that individuals try to acquire through education are generally limited. In addition, it is considerably more difficult to prepare test items that measure high-level skills than test items that measure recall. Hence, educators and institutions often repeat test items that have been used before. This may not create a problem if the students have no access to previously used items. Moreover, information on the psychometric properties of these test items will have already been obtained since the items have already been

used on a similar sample. However, the possibility that some or all of the test items have been previously used and are easily accessible to students will result in measurements focusing only on memorization, even though these items are developed to measure high-level cognitive skills. This may also greatly affect the psychometric properties of the items as well as the reliability, measurement, decision validity, and equalization of test scores (Buckendahl, Gerrow, & Pros, 2016; Wollack, Sung, & Kang, 2006; Wood, 2009). As a result, exams may lose their meaning and significance.

This problem is related to the dilemma often encountered by educators and institutions, especially in medical training of whether items and answer keys of exams should be published and previously used items should be reused. Educators or institutions can publish items and answer keys of exams with good intentions and expectations, such as establishing an open and transparent teaching environment, helping with exam preparation, increasing students' self-knowledge, and providing students with feedback about exam performance. On the other hand, although educators or institutions do not publish items and key answers, students can collaborate among themselves to memorize the items in the exam and document them after the exam. For this reason, the reuse of the items used in the exams in the new exams requires some precautions in addition to non-publication of items and key answers.

In addition, the publication of test items and answer keys generates many discussions relevant to the social, legal, educational, and psychometric aspects.

The dilemma of whether items and answers should be published and previously used items should be reused has long existed in many countries. For instance, this issue dates back to the 1970s in the United States. In 1978, the National Academy of Sciences led a discussion on the advantages and disadvantages of publishing test items in terms of the legal, social, statistical, and educational aspects, and many competitive ideas were shared. A draft law on the publication of test items after exams was voted on in 28 states between 1977 and 1983, but the law took effect in only two states (New York and California). This law is known today as the truth-in-testing law (Dorans, 2012; Florio, 1979; Greer, 1984; Messick, 1981). Similarly, in South Korea, the National Health Personnel Licensing Examination Board of Korea decided to make the documents related to national exams public in 2012 (Yang, Lee, & Park, 2018).

In Turkey, individuals formerly had access to information on exam items under the Right to Information Act (Act No. 4982, which took effect on September 10, 2003) upon individual application and within certain conditions (e.g., when the exams did not contain any information or document that could be generated as a result of separate research or analysis). However, 'law on amending certain laws and decrees law' (Law No. 6495), which was adopted on December 7, 2013, excluded items and answer keys of exams administered by the Directorate of Measurement, Selection, and Placement Center (ÖSYM) from the scope of the Right to Information Act.

Studies on whether items and answer keys of exams should be published have mostly focused on national exams. For example, Yang, Lee and Park (2018) analyzed data obtained from the United States Medical Licensing Examination (USMLE) and concluded that there was no significant difference in pass-fail rates and test and item statistics after the test items were published. Wood, Stonge, Boulais, Blackmore, and Maguire (2010) studied the data of students who took the Medical Council of Canada Evaluation Examination (MCCEE) more than once and found that the publication of items used in previous exams did not result in a significant difference in candidates' performance.

Wagner-Menghin, Preusche, and Schmidts (2013) found that repeating some of the exam items that measure recall led to an increase in only some students' scores. Gilmer (1989) concluded that the publication of test items after exams caused a 10% increase in the number of students who passed their grades.

The findings of most of these studies indicate that the publication of test items and answer keys had no significant effect on the test and item statistics. However, almost all of these studies used data from nationwide exams (e.g., USML, MCCEE, Public Personnel Selection Exam (KPSS) and Exam Of Expertise In Medicine (TUS)) organized by professional institutions such as the Student Selection and Placement Center in Turkey, the Educational Testing Service in the United States, and the Medical Council of Canada. Considering the infrastructure and the number and areas of expertise of the employees in these institutions, one can argue that these organizations can easily cope with the burden of publishing test items and answer keys. In other words, institutions such as OSYM, MCC, ETS, USML have the infrastructure to replace these items even if they publish the items and key answers.

On the other hand, the situation may be quite different for smaller-scale exams (e.g., midterms, finals, and make-up exams) carried out internally by institutions such as universities, faculties, and institutes. Because the infrastructure facilities of institutions such as universities, faculties, institutes are not as strong as those such as

OSYM, MCC, ETS, and USML.

In tests, the possibility that some or all of the test items have been previously used and are easily accessible to students may have an impact on the results of the exams. Thus, it is necessary to examine whether using previously used items and publishing test items and answer keys has an effect on the exams administered by these institutions.

To present meaningful discussions for countries and institutions concerned about the outcomes of using previously used items and publishing test items, this study aimed to examine the effect of repeating test items on students' pass-fail rates and on the psychometric properties of tests and items. The research questions are as follows:

1) Is there a relationship between repeated test items and pass-fail rates?

2) Do the psychometric properties of the repeated test items change when they are reused?

3) Do the psychometric properties of the test change when the test items are repeated?

This study uses an empirical approach to contribute to solving the dilemma of whether to use previously used items and to publish test items and answer keys, an often-debated topic, especially in intensive and comprehensive programs such as medical training.

2. Method

2.1 Research Design

This study was planned as a basic research because it aimed to examine whether the repeated use of test items in subsequent exams had significant effects on students' pass-fail rates and on the psychometric properties of items and tests (Royce, Straits, & Straits, 1993). Ethics committee approval was received for research.

2.2 Study Sample

The study consisted of 115 tests and 11,500 items used in the midterm and final exams of 3,910 students in the preclinical term at the Faculty of Medicine from 2014 to 2019. Only 10,491 items that could be transferred to the digital environment were included in the analysis. Table 1 shows the distribution of students by year and period.

Table 1. Distribution of students by year and period

	Period I	Period II	Period III
2014–2015	265	244	219
2015–2016	271	272	215
2016–2017	269	278	249
2017–2018	292	266	260
2018–2019	299	276	235
Total	1,396	1,336	1,178

The findings show that 28.57% of the items examined were repeated items. Table 2 presents the distribution of repeated items by period and year.

Table 2. Mean percentage of repeated items by year and period

	Period I		Period II		Period III	
	M	*SD*	*M*	*SD*	*M*	*SD*
2014–2015	22.38	3.11	22.10	3.63	30.51	1.30
2015–2016	27.55	3.13	26.83	5.01	34.62	3.32
2016–2017	26.42	4.42	16.98	2.16	37.55	2.02
2017–2018	26.00	3.76	18.14	1.98	39.00	2.66
2018–2019	26.12	2.52	18.14	1.31	35.66	2.59
Total	26.52	1.67	20.02	1.59	36.75	1.26

Table 2 shows that exams held in the third period of 2017–2018 had the highest percentage of repeated items (39%), while exams held in the second period of 2016–2017 had the lowest percentage of repeated items (16.98%).

2.3 Data Analysis

Data analysis included descriptive statistics related to the total test scores, item difficulty (percentage of correct

answers for item) and item discrimination values, and internal consistency (KR-20) values for reliability. The Shapiro-Wilks test was used to analyze the distribution structure, and *t* tests were used to analyze the differences between groups.

3. Results

3.1 Relationship between Repeated Test Items and Pass-fail Rates

Findings shows a significant at the .01 level and moderately positive relationship ($r= .468, p < .01$) between the percentage of repeated items and the number of students who qualified to pass their grades.

3.2 Effect of the Percentage of Repeated Items on Their Psychometric Properties

Pearson's correlation coefficient was used to examine the relationship between the percentage of repeated items in tests and the psychometric properties of the tests.

Findings shows that the percentage of repeated items and the reliability coefficient had a non-significant, weak negative relationship ($r=-.20, p>.05$) and that the standard deviation and percentage of repeated items had a negative, weak but significant at the .05 level relationship ($r=-.26, p<.05$). There was no correlation between the percentage of repeated items and other psychometric variables (Mean, Skewness and Kurtosis).

3.3 Item difficulty and Item Discrimination Values at Initial and Repeated Use of Test Items

We examined the item difficulty (p_j) and item discrimination (r_{jx}) values of the repeated test items at initial and repeated use. Table 3 presents the findings.

Table 3. Item difficulty and item discrimination values of repeated test items at initial and repeated use

	Use of items	*M*	*SD*	*t*	*p*
Item difficulty(p_j)	Initial	0.52	0.26	-7.919	.000
	Repeated	0.81	0.17		
Item discrimination(r_{jx})	Initial	0.19	0.15	2.463	.015
	Repeated	0.11	0.23		

The mean item difficulty was 0.52 ± 0.26 during the initial use of the items. When the items were reused, this value significantly increased to 0.81 ± 0.17, $t = -7.919$, $p < .001$ (Table 3).

On the other hand, the mean item discrimination value significantly decreased from 0.19 ± 0.15 during initial use of the items to 0.11 ± 0.23 when the items were reused, $t = 2.463$, $p < .05$.

4. Discussion

The behaviors that are targeted to be acquired by individuals through education mostly demonstrate 'maximum performance behaviour' feature. Therefore, students are expected to do the most and the best in unit time by pushing their limits in exams.

Most students focus on studying for and doing well on exams. Therefore, the possibility of using some or all of the items from previous exams in subsequent ones may distract students from the main learning resources, learning objectives, and class attendance, among others, and impel them to accumulate and memorize previously used test items. Hence, the skills acquired by students may not go beyond the level of recall.

This study therefore aimed to examine the effects of using the same test items in subsequent exams. The findings demonstrated that the mean rate of repetition in test items from 2014 to 2019 ranged from 16.98% to 39.00%. In many countries, the cut-off score is taken as a fixed value independent of test and item statistics and applied as 60% of the total score (Park & Yang, 2015). Thus, it would be difficult to accurately evaluate individuals based on tests with a high rate of repeated items and a fixed cut-off score.

The findings also showed that the total score variance decreased significantly as the percentage of repeated test items increased. This may be related to the fact that increasing the percentage of repeated items narrowed the total score range. In addition, the narrowing range due to the increased percentage of repeated items may adversely affect the other psychometric properties of the test (Gulliksen, 1950; Magnusson, 1967; Park & Yang, 2015).

A significant, moderately positive relationship was observed between the percentage of repeated items used in exams and the number of students who were eligible to pass their grades. These results are consistent with those of Gilmer (1989), who found that using repeated items in exams led to a 10% increase in the number of students who passed their grades. He stated that the continuous disclosure of test items and answer keys would lead to

higher pass rates independent of the test takers' performance, which would provide unfair benefits for some students.

In terms of the psychometric properties of the items, the item difficulty values obtained from initial item use were significantly lower than those obtained from reuse (M = 0.52 vs. M = 0.82, respectively). The mean item difficulty values at initial item use indicated that these items were moderately difficult. Upon repeated use of the item, the difficulty changed to "very easy." These findings are consistent with those of Angelis, Hale, and Thibodeau (1980) and Gilmer (1989). Hale et al. (1980) reported that the disclosure of items and answer keys may affect future testing, which undermines the reliability and validity of tests.

On the other hand, the findings of this study contradict those of Yang et al. (2018), Wood et al. (2010), and Stricker (1984). Yang et al. (2018) found that publishing test items did not create a significant difference in the test and item statistics of the new tests. Similarly, Wood et al. (2010) reported that publishing items used in previous exams did not make a significant difference in candidate performance. Stricker (1984) examined the Scholastic Aptitude Test (SAT) data and concluded that the disclosure of test items did not have a significant effect on the individuals' performance. The findings of the present study may differ from those of the studies cited above because this study was based not only on the disclosure of test items but also on the repeated use of the same items in subsequent tests. Yang et al. (2018) did not specify whether the published items were used in subsequent tests; they only mentioned that the study used USMLE data. Similarly, Wood et al. (2010) did not mention whether the published items were used in subsequent tests but emphasized that their study used MCCEEdata. A similar situation was found in the study by Stricker (1984) using SAT data.

Exams such as the TUS, USMLE, and MCCEE are national/international exams; organizers of such exams are expected to be able to handle the burden of publishing test items and answer keys. The institutions organizing these exams have the necessary infrastructure, budget, and personnel, among others, to add new items to their item banks in place of the published items if they choose to disclose test items and answer keys. However, this may not be the case for smaller-scale exams that institutions such as universities and institutes carry out internally. The difference between the findings might be a direct and indirect reflection of this situation.

In Turkey, items and answer keys of exams are no longer covered by the Right to Information Act. Similarly, in the United States, the Association of American Medical Colleges (AAMC), which administers the Medical College Admissions Test, has initiated a legal process requesting the cancellation of the law requiring the publication of test items and correct answers due to the burden and copyright constraints imposed by this practice (Espinoza, 1993). This suggests that institutions such as ÖSYM and AAMC may also have difficulty dealing with the burden of publishing test items and answer keys.

The literature identifies two basic qualities of measurement tools: reliability and validity. When the variable to be measured is 'maximum performance behaviour', it is unreasonable to expect the developed test to ensure reliability and validity on its own. Most testing institutions and experts in the United States and Europe have reported that publishing test items and answer keys may result in changes to the psychometric properties of the items and tests. This may prevent the reuse of these items. A significant amount of labor, attention, and financial resources will be required to add new items in place of the disclosed items, which may create biased results in various statistical processes such as the test equating (Gilmer, 1989; Park & Yang, 2015; Veerkamp & Glas, 2000). The findings of our study support this view.

We conclude that reusing items from previous tests will not create negative consequences (except that the information/behavior/skill measured by the item is outdated) when the items and answer keys have not been published or when the students do not have access to the items used in previous exams. However, when items and answer keys are disclosed, new items should be used to alleviate the negative consequences. This will be an additional burden to the relevant institutions and experts. If the exam organizers do not have the capacity to shoulder this burden, it may not be advisable to publish test items and answer keys. In addition, even if the items and answer keys are not published, students could collaborate among themselves, for instance, to memorize a few items each and combine these items in a printed document, which they could share with other students after the exam to prepare for subsequent exams. Therefore, additional measures might be needed to prevent the disclosure of items and answers.

We recommend that this study be replicated using data from exams in different institutions and programs and with different item types.

Acknowledgments

No grant or financial support has been received for the study.

References

Bazı Kanun ve Kanun Hükmünde Kararnamelerde Değişiklik Yapılmasına Dair Kanun-6495 sayılı. (Law on amending certain laws and decrees law-Law No. 6495), (2013, December 7). *Resmi Gazete* (Sayı: 28726). Retrieved from https://www.resmigazete.gov.tr/eskiler/2013/08/20130802-1.htm

Bilgi Edinme Hakkı Kanunu-4982 sayılı (Right to Information Law-Law no. 4982), (2003, September 10). *Resmi Gazete* (Sayı: 25269). Retrieved from https://www.mevzuat.gov.tr/Metin.Aspx?MevzuatKod=1.5.4982&MevzuatIliski=0

Buckendahl, C. W., Gerrow, J. D., & Pros, C. (2016). Evaluating the impact of releasing an item pool on a test's empirical characteristics. *Journal of Dental Education, 80*(10), 1253-1260.

Dorans, N. J. (2012). The contestant perspective on taking tests: Emanations from the statue within. *Educational Measurement, 31*(4), 20-37. https://doi.org/10.1111/j.1745-3992.2012.00250.x

Erkuş, A. (2006). *Sınıf öğretmenleri için ölçme ve değerlendirme: Kavramlar ve uygulamalar*. (*Assessment and evaluation for elementary school teachers: Concepts and applications).* Ankara, Turkey: Ekinoks.

Espinoza, L. G. (1993). The LSAT: Narratives and bias. *The American University Journal of Gender, Social Policy & the Law, 1*(1), 121-164.

Florio, D. H. (1979). Congress watch: Truth in testing, funding cuts, and the Department of Education. *Educational Researcher, 8*(8), 13-26. https://doi.org/10.3102/0013189X008008013

Greer, D. G. (1984). *Truth-in-testing legislation, an analysis of political and legal consequence, and prospects.* Houston, TX: Institute for Higher Education Law and Governance.

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

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley. https://doi.org/10.1037/13240-000

Angelis, P. J., Hale, G. A., & Thibodeau, L. A. (1980). *Effects of item disclosure on TOEFL performance* (ETS Research Report, Report No. RR-08). Princeton, NJ: ETS. https://doi.org/10.1002/j.2333-8504.1980.tb01231.x

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice, 41*(4), 212-218. https://doi.org/10.1207/s15430421tip4104_2

Magnusson, D. (1967). *Test theory.* Boston: Addison-Wesley.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher, 10*(9), 9-20. https://doi.org/10.3102/0013189X010009009

Park, Y. S., & Yang, E. B. (2015). Three controversies over item disclosure in medical licensure examinations. *Medical Education Online, 20*, 1-5. http://dx.doi.org/10.3402/meo.v20.28821

Royce, S., Straits, B. C., & Straits, M. M. (1993). *Approaches to social research*. New York, NY: Oxford University Press.

Stricker, L. J. (1984). Test disclosure and retest performance on the SAT. *Applied Psychological Measurement, 8*(1), 81-87. https://doi.org/10.1177/014662168400800109

Veerkamp, W. J. J., & Glas, C. A. W. (2000). Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics, 25*(4), 373-389. https://doi.org/10.3102/10769986025004373

Wagner-Menghin, M., Preusche, I., & Schmidts, M. (2013). The effects of reusing written test items: A study using the Rasch model. *ISRN Education, 15*(3), 1-7. http://dx.doi.org/10.1155/2013/585420

Wollack, J., Sung, H. J., & Kang, T. (2006). *The impact of compounding item parameter drift on ability estimation.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

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

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

Yang, E. B., Lee, M. A., & Park, Y. S. (2018). National health personnel licensing examination board of Korea. *Advances in Health Sciences Education, 23*(2), 265-274. https://doi.org/10.1007/s10459-017-9788-8