Research Article

The Effects of Reusing Written Test Items: A Study Using the Rasch Model

Michaela Wagner-Menghin,¹ Ingrid Preusche,¹ and Michael Schmidts²

¹ Department for Medical Education, Medical University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria

² Institute of Medical Education, Curriculum Coordination & Students' Office, University of Bern, Murtenstrasse 11, CH-3010 Bern, Switzerland

Correspondence should be addressed to Michaela Wagner-Menghin; michaela.wagner-menghin@meduniwien.ac.at

Received 25 February 2013; Accepted 5 April 2013

Academic Editors: K. Capps and T. Hovardas

Copyright © 2013 Michaela Wagner-Menghin et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Background. Relevant literature reports no increase in individual scores when test items are reused, but information on change in item difficulty is lacking. **Purpose.** To test an approach for quantifying the effects of reusing items on item difficulty. **Methods.** A total of 671 students sat a newly introduced exam in four testing shifts. The test forms experimentally combined published, unused, and reused items. Figures quantifying reuse effects were obtained using the Rasch model to compare item difficulties from different person samples. **Results.** The observed decrease in mean item difficulty for reused items was insignificant. Students who self-scheduled to the last test performed worse than other students did. **Conclusion.** Availability of leaked material did not translate into higher individual scores, as mastering leaked material does not guarantee transfer of knowledge to new exam items. Exam quality will not automatically deteriorate when a low ratio of randomly selected items is reused.

1. Background

A written multiple-choice assessment is cost efficient when testing large cohorts. However constructing new items each year, or for each new test to prevent their leakage and to maintain test validity and fairness, offsets the cost efficiency. Therefore, the question arises under which condition one can safely reuse items. Literature on the construction of written test items [1, 2] and cheating on written tests [3, 4] provides four conceptual factors that should be considered in the discussion of reusing written test items.

1.1. Reuse Expectation. When items are not reused, students need not pass their contents to subsequent candidates because, aside from allowing one to become familiar with the test format, there is little benefit in using the original test material for studying. However, when students expect to encounter reused items on the test, passing item content to subsequent candidates and trying to obtain passed item content prior to the exam for study purposes, may help in obtaining a good grade.

1.2. Cheating Attitude. There are cultural differences regarding the concepts of cheating and cheating behavior [4]. Teixeira and Rocha [4] found that the average magnitude of academic cheating by university students is quite high (62%). They also found differences in rates of cheating between countries (e.g., the USA: 30%–40%, Germany: 50%–60%, Austria: 70%–80%, and France: 80%–90%). Studies on cheating usually focus on the copying of tests or assignments, not on the passing of test items or use of test items for studying; however, some consider these behaviors as forms of cheating. It can be said that, in some cultures, it is common for students to regard passing exam content to fellow students and using passed content for studying not as cheating but as cooperative and strategic study behavior. One study [3] states that 23% of graduate business students and 18% of nonbusiness students in the USA and Canada admit to cheating on tests, with the business students engaging more in “learning what was on a test from a student who took it in an earlier class.” However, studies do not usually give figures on the prevalence of this type of cheating.

1.3. Exam's Consequences. Exams are given in different stages of education and with different functions, some with severe consequences in the event that the student fails. The pressure to succeed towards the end of the educational term promotes cheating behavior [4].

1.4. Item Content. Item writing guidelines classify written test items that require an examinee to reach a conclusion, make a prediction, or select a course of action as "application of knowledge items" and items that test only memory for isolated facts (without requiring their application) as "recall items" [1]. Reusing the former may be more problematic with respect to test fairness and validity. On its first use, an item may require the examinee to reason out the answer by applying basic principles, but once it has been passed to other students, it might require simply recalling the correct answer. This will not be a problem with recall items.

But what happens with item difficulty and person scores, when test items are reused? Some studies focus on the retesting of persons or reuse of written test forms. One approach uses the same test form on different occasions for different candidates to clarify whether candidates on the later occasions benefit from more study time [5, 6]. Results indicate no score difference [5] or better scores for early test takers [6]. Another approach uses the same (and a different) test version on different occasions for the same candidates to investigate whether taking a test repeatedly gives candidates an advantage [7, 8]. In general, repeat candidates achieve higher scores, but there are no further advantages for repeat candidates assigned to the same test form. The cited studies do not elaborate on whether students expected to encounter reused items, but as the studies were conducted in the USA and Canada (two in the setting of in-class assessment, two in that of credentialing), students' preparedness to share information on exam content is assumed to be low. To our knowledge, no studies have yet been conducted in settings where students expect to encounter reused items or in settings where leaking content and using leaked content for studying is not regarded as cheating.

In addition, the cited studies focus on differences in individuals' mean scores. Thus, they leave open the question of what happens to an item's level of difficulty when it is reused. However, a possible systematic leakage of items and, thus, corrupted validity of items will not be detected by a research design that focuses only on persons' mean score difference. Focusing on these effects requires quantifying differences in item difficulty between the item's first use and its reuse. However, statistics used for expressing item difficulty, in general, are sample dependent, precluding meaningful interpretation of differences in item difficulty between different person samples. Thus, to our knowledge, no studies have yet investigated the effects of reuse on item difficulty.

2. Purpose

The present study seeks to test a research design that allows quantifying the effect of reuse on a single test item's level of difficulty, independent of the individual test takers. It

addresses reuse effects with students who are accustomed to encounter a high percentage of reused items in subsequent testing shifts and who consider discussing test items not as cheating. The study uses available written items complementing a practical assessment (PA) on basic clinical procedures in undergraduate medical education.

Hypothesis 1: As the available written items target recall of facts as well as application of knowledge and passing of item content between the testing shifts is common, a decrease in item difficulty is expected for reused items, especially for those requiring application of knowledge.

Hypothesis 2: As students expect to encounter reused items and they also routinely post exam contents on Internet platforms, which are accessed by other students, it is expected that the measures of the ability of those who attempt a test later in the year (and who thus encounter reused items) will increase.

The study also seeks to measure the proportion of students who are motivated to benefit from the leakage of test content.

Hypothesis 3: Based on international literature on self-reported cheating in higher education in general and test cheating specifically, it is expected that between 30% and 60% of the students are motivated to benefit from leaked content. The literature on this hypothesis is admittedly sparse, as the only data available are from self-reports on cheating; thus, the investigation of this hypothesis is clearly explorative in character.

3. Method

During the 2009 spring term, 671 students attempted the newly introduced in-course exam that assessed basic clinical skills. To comply with students' curiousness about the items, a representative set of items (one-best-answer multiple-choice questions (MCQs)) were published in the official study materials. Four test forms that experimentally combined three types of items ("*published*" = *not new to students*, "*first use*" = *new to students*, and "*reused*" = *possibly not new due to leakage*) were administered on four days within a three-week period to investigate Hypotheses 1 and 2. Test forms 2 and 3 contain 35% reused items, and form 4 contains 45%. The procedure of assigning students to one of the four test days constitutes the field experiment to investigate Hypothesis 3. The assessment was implemented as a compulsory formative pretest to the summative exam, but to stimulate the achievement motive, a bonus was given for the summative mark when the students obtained a written grade higher than the practical grade. Students were assigned a day to take the test but were allowed to change it to accommodate their individual schedules. For a student population that is, used for preparing for exams by studying previous exam material, the least attractive day for taking a newly introduced exam is the first day because only the published sample items are available for additional studying. Thus, students who were assigned to take the exam on day 1 and kept this date and those

who voluntarily scheduled to take it on day 1 can be regarded as least interested in benefitting from leaked content (first test date). Their performance is compared with two other groups: those who had the opportunity to benefit from leaked content but did not bother to increase this opportunity (follow-up test date as scheduled) and those making an effort likely to increase their opportunity to benefit from item leakage by rescheduling to the latest possible test date (strategic reschedules). Between the test days, a popular online student platform was screened, and it was observed that item content was discussed vividly by students online.

It is acknowledged here that this sampling procedure has some drawbacks from an experimental point of view. Because of the rescheduling, there is no longer a random assignment to the four test days. Thus, the groups formed on basis of their rescheduling behavior may contain students with different knowledge and skills. As one of the reviewers pointed out, there may be several reasons for rescheduling, beside the interest to benefit from leaked content. However, the exact reason for rescheduling could not be surveyed in the given setting. These sampling issues however do not affect all our hypothesis testing in the same way. It is clear that making assumptions about motivational differences between the three groups as defined above and thus the testing of Hypothesis 3 remains speculative. On the other hand, this sampling mirrored the real life situation when more than one test day is set up and students are free to schedule for a date. As a result, this sampling makes it possible to address Hypothesis 2 in a realistic setting. Finally, as this study's main focus is on quantifying changes in items difficulty these sampling issues will not affect the testing of Hypothesis 1. As explained below the chosen psychometric framework allows objective comparison of items difficulty independent from the sample used.

To estimate item difficulty and person ability measures necessary for examining Hypotheses 1 and 2, a probabilistic psychometric framework (Rasch model) was chosen as the measurement model. The Rasch model is a mathematical formula that combines the answers a person gave to a set of items with ideas about the relative strength of persons and items in a way that *absorbs the inevitable irregularities and uncertainties of experience systematically by specifying the occurrence of an event as a probability rather than a certainty* ([9], page 4). This model was chosen for its ability to express item difficulties relative to each other. Item difficulty is estimated independently from the persons' ability and simultaneously for multiple-item/person sets with overlapping items. Thus, items are equated, and we can make an objective comparison of item difficulties estimated using data from different person samples. Persons' ability is expressed using the same scale as with item difficulty and accounts for such. Solving, for example, *nine items out of ten* gives a lower ability measure with easier items, but a higher measure for more difficult items.

To prepare the testing of Hypotheses 1 and 2, the data of the four test days and test forms are organized in two different data matrices (Figure 1). Using matrix (a), a conventional incomplete data matrix with common items to link the test forms item difficulties is estimated and used for estimating

person measures to test Hypothesis 2. Additionally, items' fit with the model is tested statistically (infit/outfit-statistics, differential item functioning), thus contributing to the construct validity of the written exam. As reliability coefficient, the reliability of person separation (rel_{ps}) is calculated; this index is based on the same concept as Cronbach's alpha (α) and is interpreted as such. Matrix (b) is used to estimate the item difficulties to address the change in level of item difficulty in reuse. In this matrix, data of the reused items are treated independently from the data of the first use. The data of 33 (first use) plus 15 (second use) plus 2 (third use) items are input in the analysis to estimate item difficulties. These item difficulties are used to evaluate difficulty changes for Hypothesis 1.

Hypothesis 1 was tested by performing a paired-sampled *t*-test wherein the difficulty levels of the first use and second use were compared. In order to explore in detail if application of knowledge items (6 items) are more prone to difficulty change than recall items (9 items), a 95% confidence interval was constructed around each item's difficulty level to compare first use with reuse difficulty level.

To test Hypothesis 2, an ANOVA was performed by comparing the ability measures of confident performers, compliant performers, and strategic performers as defined above. Hypothesis 3 was tested descriptively by deriving percentage values. The percentage of students who changed their original assignment of taking the exam on the first day serves as an estimate for the percentage of students motivated to benefit from leaked item content. To control for achievement motivation, only 535 students (80%) actually taking the exam as a pretest to the summative PA are included in evaluation of Hypotheses 1 and 2.

4. Results

4.1. Descriptive Statistics. Table 1 shows the mean and standard deviation of the item easiness parameter. Table 2 gives the mean and standard deviation of the measures of ability as well as an overview of how the number of students fluctuates across the four test days. Of the 158 students assigned to day 1, 75 (47%) stayed. Together with 20 students who opted for day 1 over a later assignment (days 2, 3, or 4), they form the subsample of students taking the "first test date" ($n = 95$), that is, those who do not seek to benefit from leaked content. Of the 366 students assigned to days 2 and 3, 200 (55%) stayed. They form the group of "students staying as assigned", that is, those who could benefit from the leaked content of previous dates but did not attempt to delay the exam further. Students who rescheduled to day 4 ($n = 60$) form the subsample of "strategic reschedulers." Their decision to choose the last day is interpreted as an attempt to increase their opportunity to benefit from leaked content. Students who were assigned to and remained at day 4 are not included in one of those groups as they could not postpone taking the test further.

4.2. Item's Psychometric Quality. Fit statistics do not indicate misfit (Table 3) or differential item functioning ($LR = 28.7$; $P = 0.43$). The mean standard error of estimation ($SEE = 0.76$

FIGURE 1: Symbolized data matrices for estimating item difficulty parameters for Hypotheses 1 and 2. Items are in the columns and persons in the lines; shaded areas contain persons' answers coded as 1 = correct, 0 = not correct; white areas contain no data, indicating that the person has not worked on the respective items. It is possible to estimate parameters using data from all four item/person groups simultaneously because there is a set of common items.

TABLE 1: Mean and standard deviation of item easiness parameter.

	M/SD	Min/Max	N
Item easiness parameter data matrix (a)	-0.06/1.17	-2.95/2.98	33
Item easiness parameter data matrix (b)	0.0012/1.33	-2.95/3.32	50
First use (items)	-0.52/1.23	-2.95/1.46	15
Second use (items)	0.13/1.68	-2.66/3.32	15

logits) is, as expected, reached with 14–19 items. The 95% CI (± 1.49) covers only 56% of the logit scale's range (5.28 logits); therefore, the ability measure derived out of the items' score does discriminate between persons. Given the samples' homogeneity ($SD = 0.94$ logits), the discrimination's impact is limited ($rel_{ps} = 0.39$). This low reliability of the ability measure is due to the low number of items used for this study. As this study's main target is the items difficulties, which are estimated on basis of a large person sample and not the person's ability measures we do not consider them a problem. Within this paper, persons' ability measures are therefore only compared on a group level and because of low reliability not on the individual level. To further illustrate, the items' quality reliability adjusted for test length is calculated, indicating that using 30 such written items will have increased rel_{ps} up to 0.80, which is commonly suggested suitable for test scores used for individual decision (Table 4).

Hypothesis 1: mean easiness parameter for the 15 items increases between first use ($M = -0.52$; $SD = 1.13$) and second use ($M = 0.13$; $S = 1.68$), indicating that items are easier when used a second time; this conforms to the hypothesis that reused items are easier. Still, the paired-sample t -test misses significance slightly ($T = -1.985$, $df = 14$; $P = 0.067$); thus, the hypothesis has to be rejected. However, the post hoc effect size is almost medium ($ES(d) = 0.41$), and the power is small

(post hoc power = 0.31) because of the small sample of items in our reuse experiment. With the effect size of 0.41, it would have been necessary to include 39 items in the reuse study to perform a test with a power of 0.80.

Comparing first-use difficulty with reuse difficulty by means of comparing the CIs showed that difficulty for three out of six applications of knowledge items (item numbers according to Table 3: 31, 33, 45, 36, 34, and 15) declined. The same is true for two of the nine recall items (item numbers according to Table 3: 25, 40, 41, 44, 47, 48, 49, 6, and 30). Examining the content of those two items carefully revealed there might be problems regarding the agreement of standards between teachers.

Hypothesis 2: the three subsamples' mean ability measures (Table 2) differ significantly (ANOVA: $P < 0.05$, small effect $f = 0.16$, power = 0.77). The post hoc Scheffé test indicates that "strategic reschedulers" perform worse than students taking the test at day 1 and those taking the test as assigned on a follow-up date (days 2 and 3). This finding does not conform to the hypothesis of an increasing mean score due to candidates strategically benefitting from reused items in later tests. The subjects taking the test last perform the worst.

Hypothesis 3: of the 158 students assigned to day 1, only 75 (47%) stayed. Thus, more than half of the students preferred not to take the new test without the opportunity to study passed item content.

5. Discussion

The present study developed a research design quantifying the effect of reuse on a single test item's level of difficulty, independent of the individual test takers. Using the Rasch model as psychometric model it was possible to estimate jointly item difficulties for first use and reuse and thus compare them meaningfully independently of the underlying

TABLE 3: Psychometric results.

I.Nb.	Item difficulty (classical test theory)				Item difficulty and fit statistics (Rasch model)					
	Test 1 <i>n</i> = 97	Test 2 <i>n</i> = 183	Test 3 <i>n</i> = 67	Test 4 <i>n</i> = 188	Number of persons per analysis			First use/reuse analysis		
					Test 1–Test 4 basic analysis			N = 535		
					Number of items per analysis			N = 535		
					19	14	14	15	33	33 + 15 + 2 = 50 items
	p	p	p	p	Easiness	Outfit	Infit	1st use	2nd use	3rd use
Published items										
1	0.99	0.98	0.94	0.97	1.66	0.68	0.86	1.86 (0.28)	—	
19	0.79	0.79	0.85	0.25	−1.63	1.07	1.04	−1.49 (0.12)	—	
23	0.96	0.95	0.91	0.88	0.56	1.02	0.94	0.75 (0.18)	—	
37	0.86	0.85	0.81	0.94	0.02	0.89	0.96	0.21 (0.16)	—	
New in test 1										
5	0.90				−0.11	0.80	0.97	−0.25 (0.36)		
7	0.91				0.01	1.13	0.96	−0.13 (0.38)		
25	0.87		0.76		−0.71	1.00	0.96	−0.55 (0.33)	−1.05 (0.36)	
27	0.98				1.60	0.52	0.88	1.46 (0.72)		
31	0.89	0.91			−0.11	0.61	0.81	−0.35 (0.35)	−0.01 (0.29)	
32	0.92				0.14	1.16	0.96	0.00 (0.39)		
33	0.94		0.99		0.95	0.76	1.00	0.32 (0.44)	2.28 (0.58)	
40	0.92		0.87		−0.05	0.96	0.97	0.00 (0.39)	−0.24 (0.42)	
41	0.82	0.78			−0.98	1.03	1.02	−0.87 (0.31)	−1.07 (0.23)	
43	0.78				−1.01	1.11	1.01	−1.14 (0.29)		
44	0.72	0.75			−1.35	0.94	0.96	−1.49 (0.27)	−1.24 (0.22)	
45	0.40	0.45	0.40		−2.74	1.02	1.02	−2.95 (0.27)	−2.66 (0.21)	−2.73 (0.32)
47	0.94			0.97	1.63	0.73	0.88	0.32 (0.44)	2.90 (0.59)	
48	0.65			0.64	−1.21	1.05	1.07	−1.84 (0.26)	−0.60 (0.20)	
49	0.98	0.98		0.96	1.88	0.71	0.89	1.46 (0.72)	1.56 (0.52)	2.61 (0.51)
New in test 2										
2		0.99			2.97	0.29	0.84	2.98 (1.00)		
6		0.90	0.93		0.03	0.86	0.96	−0.08 (0.29)	0.50 (0.52)	
8		0.86			−0.55	0.91	0.92	−0.52 (0.26)		
30		0.87		0.62	−0.80	0.82	0.91	−0.37 (0.27)	−0.65 (0.25)	
34		0.65		0.60	−1.47	1.08	1.06	−1.78 (0.21)	−0.79 (0.20)	
New in test 3										
4			0.91		0.22	0.51	0.72	0.28 (0.48)		
15			0.85	0.98	1.30	1.13	0.90	−0.38 (0.41)	3.32 (0.71)	
29			0.76		−1.10	1.01	1.03	−1.05 (0.36)		
36			0.94	0.70	−0.49	0.87	0.93	0.76 (0.56)	−0.30 (0.21)	
50			0.93		0.44	0.96	0.78	0.50 (0.52)		
New in test 4										
3				0.82	0.15	0.82	0.93	0.53 (0.24)		
9				0.84	0.29	0.73	0.90	0.67 (0.25)		
10				0.82	0.15	0.88	0.93	0.53 (0.24)		
42				0.87	0.34	0.75	0.91	0.72 (0.25)		

I.Nb. = Item number; *p* = item difficulty (classical test theory); easiness = relative item difficulty (probabilistic psychometric analysis); higher values indicate easier items; s.e. = standard error; outfit/infit = fit statistics (indicating the degree of fit with the Rasch model); values around +1 are regarded as indicating a fit with the Rasch model; values close to 0 indicate an overfit (less variation than expected); values exceeding 1 indicate an underfit (more variation than expected).

TABLE 4: Results of reliability analysis.

Day	Reliability				
	n_p^*	n_i	R	R^*	R_{30}
1-cw20	95	19 (4/-/15)	0.27 ¹	0.36	0.67 ¹
2-cw21	183	14 (4/5/5)	0.39 ¹	0.29	0.55 ¹
3-cw22	67	14 (4/5/5)	0.55 ¹	0.57	0.80 ¹
4-cw23	188	15 (4/7/4)	0.32 ¹	0.28	0.51 ¹
Total	535	33	0.39 ²	—	0.80 ²

cw: calendar week 2009; n_p^* : number of persons taking test as a pretest; n_i : number of items in analysis (published/reused/new); R : reliability; R^* : Cronbach's alpha/day without previously published items; R_{30} : reliability adjusted for test length (30 items).

¹ α : Cronbach's alpha (is listed complementary for illustrative reasons).

²rel_{ps}: reliability of person separation.

knowledge items. Repeating the study with content allowing for more application of knowledge items would contribute in further exploring the effects of reuse.

The study expands the knowledge on written item reuse in three ways. First, it provides four conceptual factors for discussing reuse effects and a methodological approach to study reuse effects on item difficulty. Considering factors (1) and (2) of the framework for discussing reuse effects presented in the introduction, it shows that even in a setting where the students expect item reuse and routinely pass exam content to subsequent candidates, reusing items in exams does not necessarily mean that the items' level of difficulty becomes easier, especially recall items' difficulties stay stable. Second, the study shows that similar to previous studies in other settings [5, 6], in our setting, reusing items in exams does not necessarily mean that persons' mean scores on the test will improve in the different tests that include reused items. This lack of benefit for those taking a latter test has previously [5, 6] been interpreted as a general deficit in study organization and time management for late test takers, which is also a plausible interpretation in our case. Given the conceptual character of the study and the resulting limitations, the tentative conclusion is that, in a written test, 30% to 45% reused items, especially when recall items are reused, will most likely not help students who seek to benefit from studying reused items.

Acknowledgments

The authors would like to acknowledge Dr. S. Meryn (Head, Department of Medical Education, Medical University of Vienna) and Dr. M. Lischka, who encouraged and supported this research. The authors would also like to thank the editors of this journal and the anonymous reviewers, whose work and suggestions improved the paper.

References

- [1] S. Case and D. B. Swanson, *Constructing Written Test Questions for the Basic and Clinical Sciences*, National Board of Medical Examiners, Philadelphia, Pa, USA, 1998.
- [2] R. Krebs, *Anleitung zur Herstellung von MC-Fragen und MC-Prüfungen für die ärztliche Ausbildung [Guidelines for the Production of MC-Items and MC-Examinations in Medical Education]*, Institut für Medizinische Lehre, Abteilung für Ausbildungs-und Examensforschung, Bern, Switzerland, 2004.
- [3] D. L. McCabe, K. D. Butterfield, and L. K. Treviño, "Academic dishonesty in graduate business programs: prevalence, causes, and proposed action," *Academy of Management Learning and Education*, vol. 5, no. 3, pp. 294–305, 2006.
- [4] A. A. C. Teixeira and M. F. Rocha, "Cheating by economics and business undergraduate students: an exploratory international assessment," *Higher Education*, vol. 59, no. 6, pp. 663–701, 2010.
- [5] C. Kreiter, M. W. Peterson, K. Ferguson, and S. Elliott, "The effects of testing in shifts on a clinical in-course computerized exam," *Medical Education*, vol. 37, no. 3, pp. 202–204, 2003.
- [6] E. R. Burns, J. E. Garrett, and G. V. Childs, "A study of student performance on self-scheduled, computer-based examinations in a medical histology course: is later better?" *Medical Teacher*, vol. 29, no. 9-10, pp. 990–992, 2007.
- [7] M. R. Raymond, S. Neustel, and D. Anderson, "Retest effects on identical and parallel forms in certification and licensure testing," *Personnel Psychology*, vol. 60, no. 2, pp. 367–396, 2007.
- [8] T. J. Wood, "The effect of reused questions on repeat examinees," *Advances in Health Sciences Education*, vol. 14, no. 4, pp. 465–473, 2009.
- [9] B. N. Wright and G. N. Masters, *Rating Scale Analysis: Rasch Measurement*, MESA Press, Chicago, Ill, USA, 1982.
- [10] J. R. Boulet, D. W. McKinley, G. P. Whelan, and R. K. Hambleton, "The effect of task exposure on repeat candidate scores in a high-stakes standardized patient assessment," *Teaching and Learning in Medicine*, vol. 15, no. 4, pp. 227–232, 2003.