Post-test analysis of automatically generated multiple choice exams: a case study

Marko Čupić[1], *Jan Šnajder*[1], *Bojana Dalbelo Bašić*[1]

Faculty of Electrical Engineering and Computing, University of Zagreb[1]

Key words: *knowledge assessment, multiple choice questions, post-test analysis, exam generation software*

Abstract:

Multiple choice exams (MCEs) are widely used to assess students' knowledge because they can be graded objectively, consistently, and fast. An aspect of MCEs often neglected is that they can also provide a valuable feedback to the teachers. MCE post-test analysis can be used to pinpoint invalid items or assess the clarity of taught topics. In this paper, we focus on the analysis of MCEs, elaborate on the kind of problems that can arise with MCEs, and how they can be detected. We describe our experiences gained on the Artificial intelligence course taught at our Faculty, and discuss the lessons learned.

1 Introduction

Multiple choice exams (MCEs) are often used for both formative and summative knowledge assessment, especially for large enrolment classes. Although the effectiveness of MCEs is still being debated, evidence has been gathered in support of it [1,2,3,4]. If adequately designed, MCEs can test what may be considered higher levels of cognition according to Blooms taxonomy, rather than just simple recall of facts [5]. From a practical point of view, the obvious benefits of MCEs include objective, consistent and fast, possibly automatic grading. The latter is important for formative assessments for which fast feedback to the students is essential. It is, however, often overlooked that MCEs can also provide a valuable feedback to the teachers. With appropriate post-test analysis, the clarity and comprehensibility of certain topics can be assessed, while at the same time the quality of MC questions can be systematically measured and improved upon.

In this paper, we focus on the post-test analysis aspect of MCEs and describe our experiences gained on the Artificial intelligence course taught at our Faculty. On this course, taught to about 150 students, MCEs were used for summative as well as for short formative exams. We used the previously developed *Enthusiast* tool to generate the exams automatically [6]: the MC items were polled from a database of questions and their presentation order was randomized. For analysis of test results, we use *Ferko*,[1] a new and promising course management system developed at our Faculty. These tools provide a convenient framework for generation and analysis of MCEs.

The purpose of our post-test analysis is threefold. First and foremost, we evaluate the validity and usefulness of MC items with the aim of improving the effectiveness of subsequent MCEs.

1 https://ferko.fer.hr/ferko

To this end, we measure the discrimination index and the difficulty indices, and discuss how these can be used within our framework to pinpoint erroneous or methodically flawed items. Secondly, we take a look at how topic clarity and comprehensibility can be estimated within our framework. Finally, we investigate how the choice and the arrangement of items affect the MCE effectiveness

The rest of the paper is structured as follows. In the next section we describe the methodology used in this work: the MCE framework and the means for MCE post-test analysis. In Section 3 we apply this methodology in our case study. Section 4 concludes the paper and outlines future work.

2 Methodology

2.1 MCE generation

In this work we focus on paper-and-pencil MCEs generated automatically using the Enthusiast tool [6]. Enthusiast takes as input a plain-text database of single-response MC items and a test specification provided by the user, and generates as output randomized MCEs (i.e., MCEs that differ among themselves to some extent). In order to improve the variability across test sheets, each MC item in the database may be split into several item variants that refer to the same topic, but differ in parameters or wording. An example of a MC item from our *Artificial Intelligence* course database is given in Fig. 1. To further improve variability, each MC item (or item variant) in the database may have a redundant number of keys or distractors (i.e., the number of answer options in the database may be greater than the number of options that is actually presented to the student). As shown in Fig. 1, an MC item may be tagged with user-defined tags; based on these tags, a test specification can be given defining the content and type of the test.

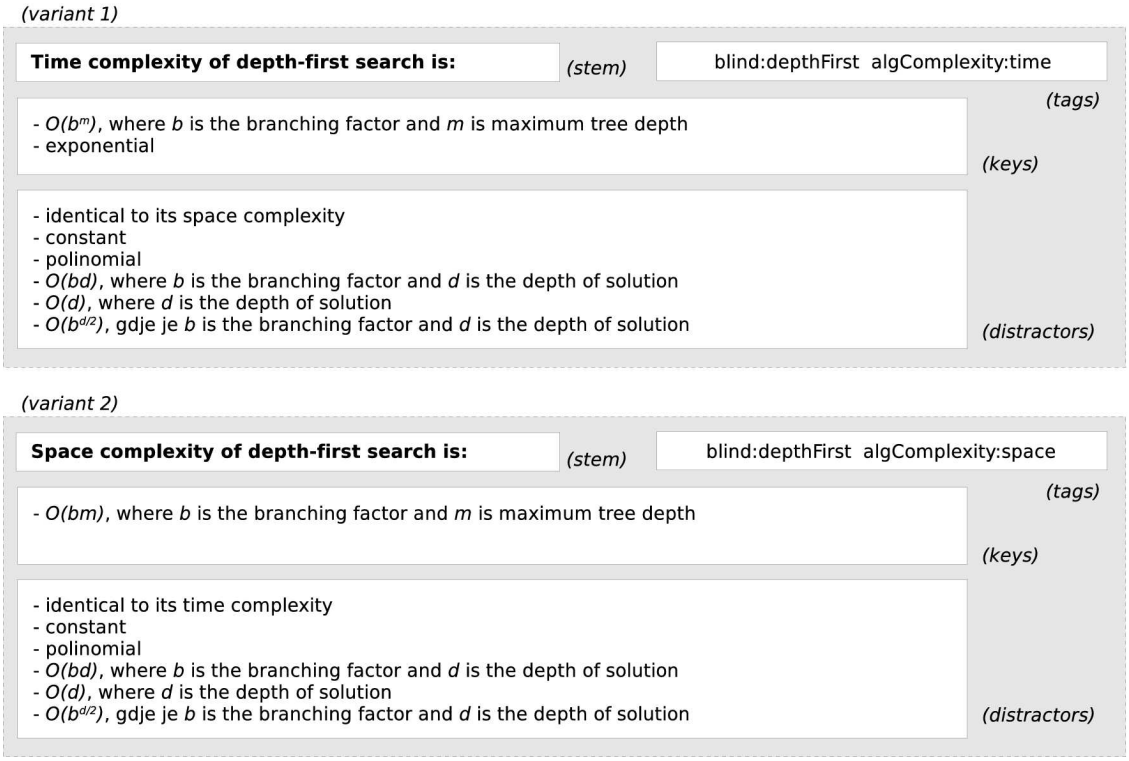

Figure 1. An MC item with two item variants.

For the Artificial Intelligence course, we have compiled a database of over 300 single-response MC items and over 550 item variants. The items are tagged with tags that relate the items to course topics. To prevent test cheating, we use Enthusiast to generate randomized test sheets with items from predefined topics. For summative exams, we create different test groups by (i) selecting at random a variant of a predefined item, (ii) selecting at random a set of answer options, and (iii) shuffling the order of items and answer options. The three summative exams (two midterm exams and one final exams) consisted of 15, 20, and 25 MC items, respectively, each with six answer options (one key and five distractors). The formative exams, given in the form of end-of-lecture quizzes, are more challenging because they are administered during lecture classes in a much less controlled setting. Thus, for end-of-lecture quizzes, we randomly vary not only the item variants, but also the items themselves (restricted, of course, to items from predefined topics). Eight end-of-lecture quizzes were given, each containing six MC items with four answer options.

2.2 MCE post-test analysis

Statistical processing is performed on MC item and MC item variants that can not be answered partially – they are either scored as correct or incorrect. They are multiple choice questions with single correct option.

For each item and for each item variant, we calculate (i) the discrimination index [7], (ii) absolute difficulty, (iii) relative difficulty, (iv) number of students that received the item, (v) number of students that answered the item correctly, (vi) number of students that answered incorrectly, and (vii) number of students that did not answer the item (blank answer).

Formulas used for calculation of (i) discrimination index DI, (ii) absolute difficulty AD, (iii) relative difficulty RD are as follows. To calculate the discrimination index, students are ranked by the total MCE score. From this list, two groups of students are considered: the upper 25 percent (U) and the lower 25 percent (L) of students. Given an item *i*, for each of the two groups we calculate the group score obtained on that item (denoted by $score^{U}_{i}$ and $score^{L}_{i}$) and the maximum possible score for that item (denoted by $possibleScore^{U}_{i}$ and $possibleScore^{L}_{i}$). Discrimination index of item *i*, denoted DI_i , is then calculated as follows:

$$DI_i = \frac{score_i^U}{possibleScore_i^U} - \frac{score_i^L}{possibleScore_i^L} .$$

Discrimination index for item variants is calculated in the same manner. Parameter DI indicates the difference in success of answering an item that exists between the "good" and the "bad" students. Here, "good" students are taken to be those that in total achieved the best assessment results, whereas "bad" students are those that in total achieved the worst assessment results. In other words, parameter DI tells us how good an item is in discriminating between "good" and "bad" students. DI takes on values from -1 to +1. Value of zero means that all of the students from upper and lower groups were equally successful in answering the item. This is something that usually requires further investigation (to start with, we might investigate whether "equally successful" means that no one has answered that item correctly or that all students have answered it correctly). Also, it should be noted that there are alternative definitions of DI. Some authors suggest using upper and lower third of population, upper and lower 27% of population, or similar [7].

To calculate the absolute difficulty, for each item (or item variant) we count how many students received that particular item an how many answered the item correctly, and then use:

$$AD_i = 1 - \frac{students_{correct_i}}{students_{correct_i} + students_{incorrect_i} + students_{blank_i}} .$$

The absolute difficulty can range from 0 to 1, where 0 means that all students answered the item correctly, and 1 that no one answered it correctly. It should be noted that there also exists a similar measure called "Item difficulty" ID, which equals to:

$$ID_i = \frac{students_{correct_i}}{students_{correct_i} + students_{incorrect_i} + students_{blank_i}} = 1 - AD_i .$$

We chose to use AD in this work since its interpretation is more intuitive: the closer the value to 1, the more difficult the item.

Relative difficulty is calculated similarly, the only difference being that we only consider the students who have answered the item (i.e., blank answers are not treated as incorrect answers):

$$RD_i = 1 - \frac{students_{correct_i}}{students_{correct_i} + students_{incorrect_i}} .$$

Relative difficulty can range from 0 to 1: the value of 0 means that all students who answered the item, answered it correctly, whereas 1 means that no one who answered the item, managed to answer it correctly. It is easy to show that there holds $AD_i \geq RD_i$.

The case study that we turn to next is based on the above-described parameters. Before this, two points are worth mentioning. Firstly, we will conduct our analysis at both the level of items and the level of item variants. For single-variant items, there is no difference between these two. For many-variant items, the item parameters are calculated using their item variants. Secondly, as described in Section 2.1, when MCEs are generated, two otherwise identical items (or item variants) may differ in the set of answer options. Although such items may be considered somewhat different, the difference is less prominent in this case and we shall ignore it in our analysis.

3 Case study

3.1 Item analysis

In this paper we analyze the MCEs from the course on Artificial Intelligence given in academic year 2008/2009. We restrict our analysis to the second midterm exam, taken by 133 students, and six (out of eight) end-of-the-lecture quizzes, taken by on average 104 students. The total number of MC items thus covered is 140 items (20 items from midterm exam and 120 items from quizzes) and 168 item variants (35 from midterm exam and 133 from quizzes). This amounts to 1.2 variants per item (a higher item-to-variant ratio would take more effort, but this is certainly what we are aiming at).

For all of the items and item variants we calculated the above-mentioned post-test analysis parameters. Of particular interest for further analysis is the relationship between the discrimination index DI and absolute difficulty AD, depicted by the scatter plot in Fig. 2. The dashed lines delineate the area of possible DI-AD values, under the assumption that DI is calculated with upper and lower groups of 25% (see Section 2.2).

As can be seen on Fig. 2, most MC items are of low-to-moderate absolute difficulty (AD<0.5) and low-to-moderate discrimination index (DI<0.5). The distribution of MC items shown here can be well approximated by a dome-shaped curve, as reported in [8]. Starting from low AD values, the DI tends to increase, which is to be expected since more difficult items are better in discriminating between „good“ and „bad“ students. Furthermore, we can observe that most items of moderate AD values have moderate-to-high DI values. With further increase of AD, the DI tends to decrease. Such high-AD-low-DI items may be problematic for a number of reasons, as we shall elaborate below.

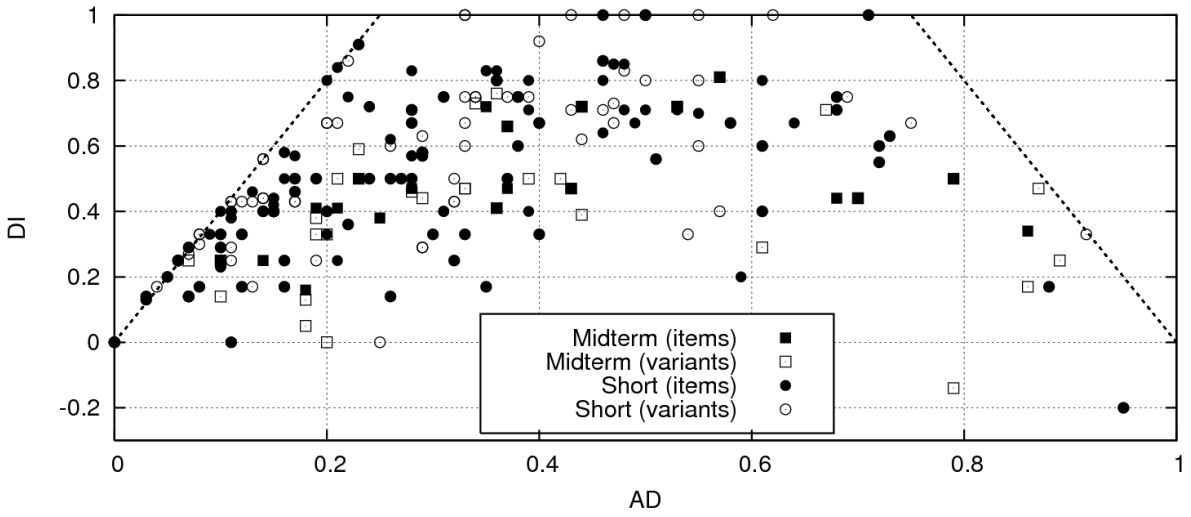

Figure 2. Absolute difficulty vs. discrimination index for midterm exams and short tests (140 items and 168 item variants)

Relying on the above insights, we set to explore four issues: the usefulness of MC items, their validity and topic clarity, as well as the effect of the order and choice of MC items.

3.2 Usefulness of MC items

We first focus on items having low discrimination index and low absolute difficulty. These are the items that were answered correctly by the majority of students. Assuming that these items are valid (i.e., that there is nothing wrong with the stem, the key, nor the distractors), we might reasonably question the usefulness of such items. So, do such items serve any purpose? The answer is no – and yes. If we look at the MCEs as a mean to produce students grades (i.e., to distinguish among "good" and "bad" students), then such items are indeed useless. However, it must be considered that some types of assessments (e.g., midterm exams, end-of-lecture quizzes, etc.) serve an additional purpose of providing a feedback to the students: how well is she or he prepared for the course and exams to come? In such cases is important not to discourage the students, otherwise they may give up from the course early in the semester. Thus, rather than being considered useless, the low-DI-low-AD items can be thought of as a kind of motivators.

It should be noted that low DI and low AD values may be an indication of nonfunctional distractors. If an item contains nonfunctional distractors, the key can be recognized by the students more easily by eliminating the nonfunctional distractors. This kind of problem can easily be detected by performing distractor analysis, i.e. by keeping track of how often distractors were chosen by the students.

3.3 Validity of MC items

To discuss the validity of the items, we will now focus on the opposite end of the AD scale: the low-ID-high-AD items. Such an item is a poor discriminator – "good" students and "bad" students answer it equally successful. However, since AD value is high, either many students answered the item incorrectly, or many students did not even try to answer it. This suggest that there is something wrong with the item, i.e., that it is invalid. Item invalidity may have to do with its stem, the key, or the distractors, and may be due to three different causes (Fig. 3). Firstly, the stem can be inappropriate. We consider a stem (or the keys) to be inappropriate if it is about a topic that was not taught well, taught differently, or even not taught at all. In general it has to do with a mismatch between how the subject was taught and how it was assessed. This might be a terminological mismatch (e.g., on lectures, a term "heuristic search" was used, whereas on MCE the synonymous term "informed search" was used), a procedural mismatch (e.g., on lectures it was taught that *step(0)=1*, whereas the item expects the students to use *step(0)=0*), or some other kind of mismatch. The second cause for item invalidity is when the stem, the distractors, or the key are ambiguous. This often happens when using negations, double negations, constructs giving raise to anaphora ambiguities, etc. Finally, the third cause for item invalidity is when the stem, the distractors, or the keys are simply – erroneous. If we consider item reusability, invalid questions should be analyzed and corrected, in order to prepare them for reusing.

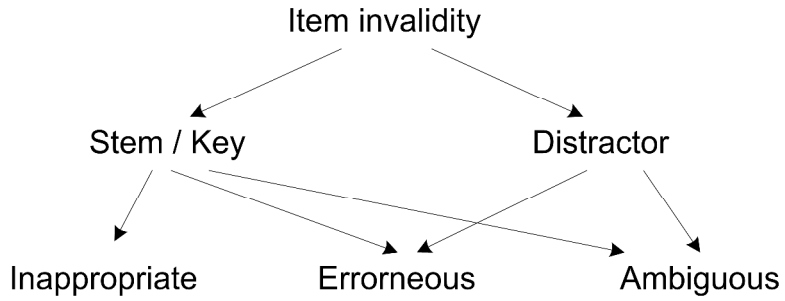

Figure 3. Taxonomy of causes for item invalidity

However, not all of items characterized by low DI and high AD are invalid. For example, an item can be perfectly valid, but demotivating. Demotivating items are usually items that are highly scored and consequently also highly penalized. In such a case, many "good" students will be unwilling to take the risk to be penalized, and they may choose not to answer the item. On the other hand, "bad" students will be more willing to risk and potentially get a high score. Consequently, the value of DI will be close to zero. Furthermore, because among the few students that answered the item most of them did not answer it correctly, AD value will also be low. Demotivating items can be detected by observing the ratio RD/AD:

$$\frac{RD}{AD} = 1 + \frac{students_{blank}}{students_{correct} + students_{incorrect}}$$

For demotivating items this ratio will be rather high. It should be noted that if grades are produced by Gaussian distribution (as it is often the case at our Faculty), these items usually will not have any effect on grades, so they should be avoided.

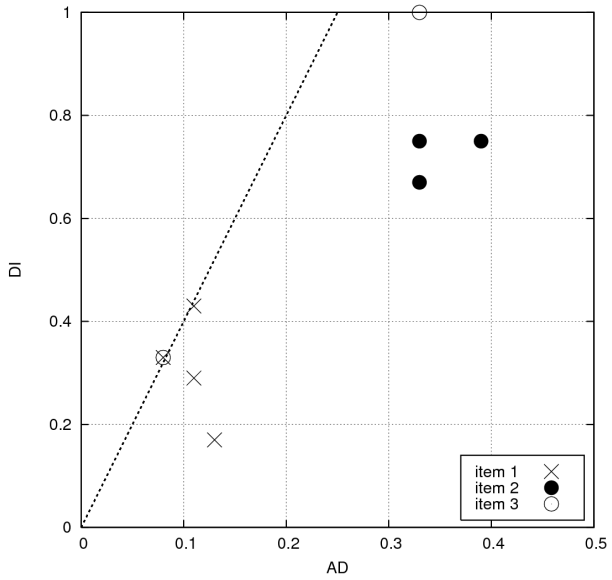

Figure 4. Scattering of variants of three MC items.

If an item has more than one variant (as in the typical case), in order to ensure fairness we would like the variants to be as homogenous as possible, i.e., roughly equally difficult and roughly equally discriminant. Unfortunately, this may not always be the case. Additional variant analysis should be performed to see if there are any outliers, and to determine the underlying causes. An example of such an analysis is given in Fig. 4. Homogeneous item variants form compact clusters in the DI-AD scatter plot, as exemplified by item 2. Item 1 is less homogeneous, calling perhaps for further investigation. This is even more the case for item 3, because its two variants very much differ in both DI and AD. A closer inspection of this particular item revealed that in this case the variants were indeed of different difficulty. The variant that came out as easier (lower AD) asked the students whether the first order logic formula $\forall x\ P(x) \vee \neg P(x)$ is tautological, satisfiable, contradiction, interpretable, or ill-formed. The variant that came out as considerably more difficult (higher AD) asked the same question for formula $\forall x\ P(x)$. First variant was answered more successfully, probably because the formula was recognized by the students as the well-known "rule of excluded middle", and therefore merely tested students' recall. The formula in the second variant, although somewhat simpler, is less typical and required some level of understanding, thus only the "good" students were able to answer it (as indicated by DI=1). In this case it may be best not to treat these two variants as variants of the same item, but rather to split them into separate items.

In general, if heterogeneous variants of an item are detected, it can also be the case that the outlier is in fact conceptually different from other variants. For example, if in one variant deals with some special case that is not dealt with by the other variants (e.g., a case in which one variant asks the student to perform a heuristic search on a graph with cycles, whereas other variants do not, which may be conceptually different) . Because of that special case, variant can be much more difficult (or much more easy) than the others, and such a variant should be promoted to a new item.

3.4 Assessing topic clarity

After all of the invalid items have been removed, remaining items can be used to assess topic clarity and provide valuable feedback to the teachers. For this we focus on the remaining items that have a medium-to-low ID and a medium-to-high AD, since these may be the items

pertaining to the topics that the students did not master well. The underlying problem can be in the low-quality or inadequate course materials, too little time dedicated to that topics, etc. Note that the prior removal of invalid items is important here because we want to measure the lack of students' knowledge rather than MCE flaws.

To perform such an analysis, items should be grouped according to topics. In our case, that can be done easily, since all items are tagged (see Fig 1), and groups can be formed simply by selecting items with specific tags. Based on items within a single group, the group's DI and AD can be calculated and then analyzed (e.g., the relation with other topics).

3.5 Choice and arrangement of MC items

In order to prevent test cheating – particularly in the case of end-of-lecture quizzes where students sit close to each other – we prefer having the order of MC items shuffled, as explained in Section 2.1. However, shuffling the order of MC items raises the question of fairness, i.e., one can reasonably ask whether the item presentation order has an influence on students' scores. We conducted a poll on 87 students, asking them whether they think that ordering by difficulty (question Q1) and by the order how topics were taught in the lectures (question Q2) would help them to obtain a better score. Results of the poll are summarized in Table 1. Almost 2/3 of the students expected that ordering by the level of difficulty would be helpful, whereas ordering by topics was perceived as helpful by only half of the students.

Table 1. Results of the poll on the order of MC items (N=87).

Poll question	yes	no	blank
Q1 – ordering by difficulty would be helpful	60.92%	37.93	1.15%
Q2 – ordering by topics would be helpful	50.57%	47.13%	2.30%

To determine whether random order of MC items indeed affects students' scores, we conducted an experiment on the final written exam. The 131 students taking the exam were randomly split into two groups: a control group consisting of 98 students that were given the exams with MC items in random order, and a test group consisting of 33 students that were given the exams with items ordered by perceived level of difficulty (easier first). Statistical analysis revealed that the average score in the control group was 13.01 ± 4.92, while in the test group it was 11.62 ± 4.93. This difference is not statistically significant at the $p=0.05$ level. Thus we conclude that the order of items does not affect the students' score and that fairness of MCE is not jeopardized by shuffling the order of MC items.

To check how good is our estimation of perceived item difficulty, which we used for item presentation ordering, and obtained item difficulty in the test group, we compared two ranks using Spearman's rank correlation coefficient, which yields a significant correlation coefficient of $\rho=0.6$ ($p=0.01$). Comparing the same ranks by Kendal Tau rank[2] correlation coefficient, we got $\tau=0.4526$ (with 2-sided p-value 0.0058). Those parameters show us that our perceived item ordering was not perfect, but it can be considered as satisfactory.

2 Kendal Tau coeficient and scatter plot was obtained using Wessa, (2008), Kendall tau Rank Correlation (v1.0.10) in Free Statistics Software (v1.1.23-r4), Office for Research Development and Education, URL http://www.wessa.net/rwasp_kendall.wasp/

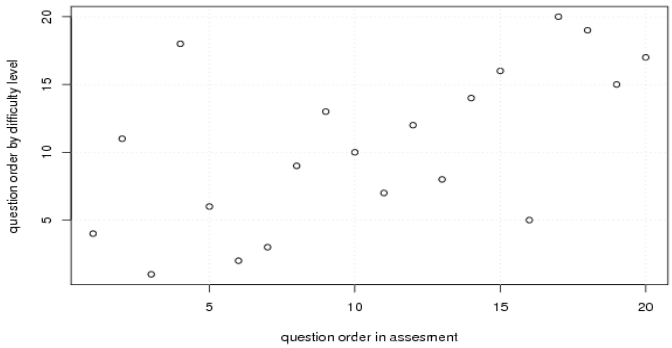

Figure 5. Scatter-plot of ranks

As for the choice of MC items, if we decide for new assessment to reuse existing items for which their DI and AD are known, then there are several issues to consider. Should we choose only items having modest (about 0.5) AD and DI values? What if we chose only items having high DI? Will it have an effect on the number of students that give-up the course early in the semester? Also, do the characteristics of an item (its DI and AD values, and possibly other indices) change over time if we include them in subsequent exams? All of these issues are to be considered as future work.

4 Conclusion

As a base for this paper we have used the actual data gathered from formative and summative assessments given on the Artificial Intelligence course at our Faculty. Our analysis suggests that many problems with MC items can be detected using parameters such as the discrimination index, absolute difficulty, and relative difficulty. If items are to be reused in subsequent MCEs, invalid items should be carefully analyzed, the cause for their invalidity should be determined, and items should be corrected accordingly. The results of our experiment on item ordering also suggest that item shuffling does not negatively affect the students' score. Thus, using shuffling as a means for cheating prevention can be recommended.

There are still many issues requiring further investigation. For example, can other item ordering strategies improve students score? Within the described framework, in order to further improve the variability of MCEs, for each item a number of variants should be created and polished, and then reused for several years. It would then be interesting to observe if DI and AD parameters of such reused items would change over time, and if so, for what reasons. Also, when considering the fairness of multi-variant items, is there an efficient technique that can be used to reliably and automatically detect if the item variants are not homogeneous? We plan to address some of these issues as a part of future work.

References

[1] Haladyna, T.M.: Developing and Validating Multiple-choice Test Items, 3rd edition. Mahwah, New Jersey: Lawrence Erlbaum Associates, 2004.

[2] Birnbaum, L.: Validity of Multiple-Choice Exam Questions. *Journal of Professional Exercise Physiology*, 6(4), 2008.

[3] Scharf, Eric M.: Assessing multiple choice question (MCQ) tests - a mathematical perspective, Active Learning in Higher Education, Vol. 8, No. 1, 31-47, 2007.

[4] Cheung D;Bucat, R.: How Can We Construct Good Multiple-Choice Items? Science and Technology Education Conference. Hong Kong, 2002.

[5] Woodford, K.; Bancroft, P.: Multiple choice questions not considered harmful. ACE 2005. Australian Computer Society, 2005.

[6] Šnajder, J.; Čupić, M.; Dalbelo Bašić, B. Enthusiast: An authoring tool for automatic generation of paper-and-pencil multiple-choice tests. Proceedings of ICL 2008, Villach, 2008.

[7] Kelly T.L.: The selection of upper and lower groups for the validation of test items. Journal of Educational Psychology, Vol. 30, 17-24, 1939.

[8] Si-Mui Sim, Raja Isaiah Rasiah. Relationship between item difficulty and discrimination indices in true/false-type Multiple Choice Questions of a Para-clinical Multidisciplinary paper. Ann. Acad. Med. Singapore. Vol. 35, 67-71, 2006.

Author(s):

Marko Čupić, mr.sc.
Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, 10000 Zagreb, Croatia
Marko.Cupic@fer.hr

Jan Šnajder, mr.sc.
Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, 10000 Zagreb, Croatia
Jan.Snajder@fer.hr

Bojana Dalbelo Bašić, prof.dr.sc.
Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, 10000 Zagreb, Croatia
Bojana.Dalbelo@fer.hr