Improving the Quality of Multiple-Choice Exams by Providing Feedback from Item Analysis

Helena Bukvova[1], Kathrin Figl[1], and Gustaf Neumann[1]

[1] Vienna University of Economics and Business, {helena.lovasz-bukvova, kathrin.figl, gustaf.neumann}@wu.ac.at

Abstract

Many universities use computer-assisted multiple-choice exams to handle large numbers of students in introductory courses. This paper describes how test-theoretical analysis can be used to offer continual feedback to examiners to ensure quality of multiple-choice exams. We present the design of an online system that gives examiners individual feedback on the test-theoretical quality of their multiple-choice exams in textual, tabular and graphical form, guiding them in item-revision and re-composition of exams if necessary.

1 Introduction

Many higher-education institutions face the challenge of assessing large numbers of students with only limited personal resources for scoring exams. One option to increase the efficiency of assessment is the use of standardised, structured exams that can be graded automatically. Among the most popular standardised assessment tools are multiple-choice (MC) exams. MC exams are a form of standardised, closed-question assessment that requires examinees to choose one or more correct answers from a collection of possible options. The type of MC question influences both their application in an exam as well as their grading (Haladyna 1992; Litzenberger et al. 2007). In its simplest form, an MC exercise consists of an item stem (e.g. a question or a problem description) and a set of answers presented to the examinees, instructing them to choose one answer that is true, whereas all others are false (distractors); this format is sometimes called single-choice question. It is also possible to include several correct options in the answer set, either informing the examinees about the number of correct options or allowing them to mark any number of options as correct (or even allow a none-of-the-above option). A slightly different format of MC questions also presents the examinees with several correct answers, but requires them to choose only one option that fits best. Further formats may require the examinees to indicate their level of confidence regarding the answer (Kolbitsch et al. 2008) or provide a brief explanation of their choice.

Advantages as the high level of standardisation and thus simplified (even automatic) grading (Chang et al. 2007) have led to a widespread use of MC-based assessment. Furthermore, the limited set of options allows MC questions to be answered comparatively quickly, thus allowing the examiners to assess a large number of areas within a limited time, providing high reliability of the results (Sim and Rasiah 2006). Some studies also suggest that MC exams can be used to test higher-level knowledge (Haladyna 1992; Kastner and Stangl 2011). On the other hand, MC exams have been criticised with regard to their validity and fairness (McCoubrie 2004); more specifically, the evaluation of partial knowledge (Chang et al. 2007) and guessing are problematic issues. To address such shortcoming, prior research has developed approaches to identify or discourage guessing (Bereby-Meyer et al. 2002), mainly through grading algorithms (Bush 1999; Lesage et al. 2013).

While exam structure and grading approaches do influence the performance of MC exams, the quality of an MC exam is predominantly determined by the wording of the items (Cheung and Bucat 2002). The importance of suitably worded MC items leads to the fact that the quality of an MC exam strongly depends on the examiners' MC- writing abilities. Although examiners can be guided and supported by guidelines, writing MC items still remains a complex and creative activity (Rodriguez 2005). In order to improve their MC-writing skills, examiners need continuous feedback on the actual performance of their MC items (Möltner et al. 2006; Sim and Rasiah 2006). Such feedback would also be valuable to higher-education institutions that reuse the pools of MC items (McCoubrie 2004; Litzenberger et al. 2007), as it would help them to identify MC items, which are suitable for further reuse or which should further be improved (McCoubrie 2004; Dickinson 2013).

To provide direct feedback on the quality of their MC items, we suggest that suitable performance indicators should be automatically calculated and made available to examiners via a learning management system. This paper describes our first insights from an ongoing project seeking to design such integrated examiner-oriented MC-feedback at the Vienna University of Economics and Business (WU). The research project can be characterized as design-oriented. Building on existing test-theoretical indicators, the main aim of the project is to create a supportive information system for examiners. In the following, we first discuss test-theoretical indicators that are useful to assess the quality of MC items and provide examiners with actionable feedback (see section 2). We then suggest a system that implements such feedback mechanisms directly into a learning management system (LMS) and demonstrate the use of the indicators with sample items (see section 3). Finally, in section 4, we discuss opportunities for further research in this context.

2 Test-theoretical analysis of MC indicators

In this section, we discuss the needs of MC examiners and give an overview on different facets of quality of MC items. We have considered the following criteria to describe a suitable MC item: (1) the item stem and answers must be understood by the examiner and the examinee in the closest possible manner, (2) the item must reflect the set, overall level of difficulty of the course, and (3) the item must display a sufficient level of challenge to allow the differentiation between different levels of knowledge among the examinees. There are two well-known and established standard indicators from classic test theory, which allow for an uncomplicated analysis of the item performance and provide examiners with clear, actionable feedback: item difficulty and discrimination index.

The *item difficulty* gives the percentage of students who have answered the item correctly (Bortz and Döring 2006). Thus, the term "item difficulty" might be confusing at first, because the higher the ex-post calculated indicator (the percentage of correct answers) is, the easier the item was. The item difficulty provides the examiners with a feedback on the performance of students on each item. Furthermore, this indicator helps to identify potentially falsely coded items, or topics insufficiently covered in the learning materials. If an MC exercise has more than one answer that is true, the “correctness” of a MC task is not dichotomous. Therefore, besides providing the percentage of students who have answered the entire question (all answers) correctly, it is also useful to indicate the difficulty of each answer option. There are no exact boundaries for optimal item difficulty. Generally, items between 35% and 80% of difficulty are considered suitable (Sim and Rasiah 2006; Venter 2010).

The *discrimination index* is based on a division of the examinees into three groups: strong (top 27% examinees), weak (bottom 27% of examinees), and middle (Möltner et al. 2006). The discrimination index describes the difference of the relative number of correct answers of the strong and the weak group. The index helps examiners to identify items that are ambiguously worded or guessable, hence not suitable to differentiate between high and low performing examinees. Again, because students can have partially correct answers, it is also useful to provide the discrimination index for each single answer option as well as for an item overall. Test-theoretical literature suggests that the discrimination index should be above 0.2; items with a negative discrimination index lead to a lower validity of the test.

More indicators have been described in the literature, but we consider those indicators presented above to be the most effective ones in providing actionable feedback to examiners.

3 Design proposal for including MC item analysis in a learning management system

While test-theoretical analysis for multiple-choice items is well established from a theoretical point of view, there are few learning management systems which support its use for instructors. Although test-theoretical analysis is used in research or to evaluate the performance of key exams such as PISA (e.g. Litzenberger et al. 2007; Venter 2010), it is rarely available in day-to-day work of examiners (with exceptions, such as Wentzel 2006). Unlike constructed-response exams, MC exams are mostly used for large student groups in which performance problems of items (such as unclear wording or unsuitable difficulty) cannot be easily adjusted. Piloting is not an option for many MC exams due to time constraints and confidentiality issues. Frequent feedback and reflection are therefore crucial for quick improvement of MC-writing skills of examiners.

3.1 Technical implementation

Institutions that use MC exams to regularly assess large groups of students typically store and manage the MC items as well as the results (students’ performance on the items in actual exams) digitally. This makes it possible to use the results to calculate item performance indicators. Figure 1 presents a diagram to illustrate how a learning management system can support the use of MC items in a teaching and learning setting based on a pool of MC items. Examiners add items to an MC-item pool and use the pool to generate new MC exams. Parts of the item pool can be opened to students to practise and prepare for exams. Students’ exam results are connected to the MC items in the pool to calculate test-theoretical indicators for the MC items. The test-theoretical results are available for examiners (and can be made available for students) through a learning management system. The item pool is then updated with new information on the quality of items, providing feedback to the examiners and enhancing the quality of subsequent test, in which items are reused.

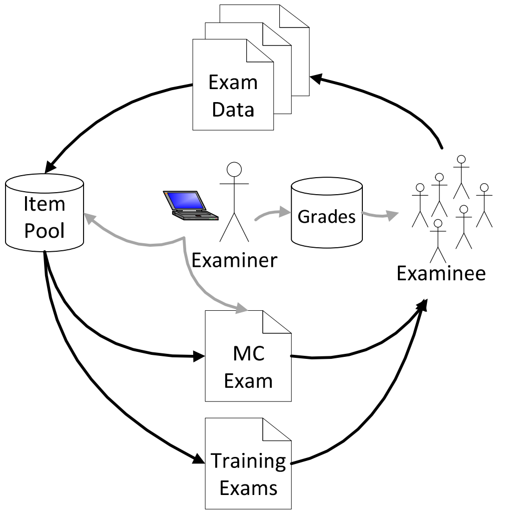

Figure 1. Learning management system to support MC exams

3.2 Exemplary application

In the following, we demonstrate the use of the indicators for exemplary items. The example items are taken from an MC exam assessing an introductory course of Business Information Systems at the Vienna University of Economics and Business. The exam contained 27 items and was taken by 451 students. Each MC item contained five answer options out of which between one and five were true (correct options are marked in the following tables with (+), distractors with (-)). The students were rewarded for marking correct answer options, and penalised for distractors. Thus, thus partial knowledge was rewarding and guessing discouraged. The MC exams were carried out in a written format and then scanned to allow for automated grading.

	Answer options					
Performance group	**a (+)**	**b (-)**	**c (-)**	**d (+)**	**e (-)**	**Overall**
Strong examinees (27%)	0.93	1.00	0.96	1.00	1.00	0.98
Middle examinees (46%)	0.84	0.99	0.78	0.94	0.99	0.91
Weak examinees (27%)	0.58	0.88	0.58	0.70	0.81	0.71
Item difficulty	0.79	0.96	0.77	0.89	0.95	0.87
Discrimination index	0.36	0.12	0.39	0.30	0.19	0.27

Table 1: Example item A

Table 1 shows the item difficulties of one example item for the three performance groups at the overall item and answer option level as well as the discrimination index (last line). The item difficulty indicators suggest that the item contains options that are too easy and thus, differentiate insufficiently between examinees. These are, in particular, the answer options *b* and *e*. These two distractors are identified as wrong by nearly all examinees. There are several possible explanations for such a result. Judging from the content of the specific item, it is likely that distractors could have been identified as wrong by most students based on common sense. The examiner would do well to replace or adapt the distractors. On the other hand, when correct options reach a high value for the item difficulty indicator, another possible explanation might be the fact that the learning goals concerning the item were achieved by all students.

	Answer options					
Performance group	**a (+)**	**b (-)**	**c (-)**	**d (+)**	**e (+)**	**Overall**
Strong examinees (27%)	0.38	0.51	0.33	0.81	0.71	0.55
Middle examinees (46%)	0.45	0.41	0.27	0.61	0.56	0.46
Weak examinees (27%)	0.56	0.53	0.23	0.54	0.51	0.47
Item difficulty	0.46	0.47	0.27	0.64	0.58	0.48
Discrimination index	**-0.18**	**-0.02**	0.10	0.27	0.20	0.08

Table 2: Example item B

Viewing the indicators in Table 2, it appears that two answer options (a, b) in this item can be solved by weak examinees better than by students with higher overall domain knowledge. This is not observable from the item difficulties, which mostly fits the 35%-80% range, but from the discrimination indices. The discrimination indices are too low for most answer options as well as overall; it is even negative for answer options *a* and *b*. The fact that weak examinees performed better than strong examinees on these two answer options can for instance indicate excessive guessing behaviour, or can be due to incorrect coding of the exercise. Since prior research has shown that weak examinees tend to guess more often than strong examinees (Sim and Rasiah 2006), this group might have an advantage with very difficult or unclear answer options. Another interpretation could be that very well prepared students realized potentially contradictory nuances of an item wording, while less prepared students just chose the answer that sounded right without reflecting it in detail. In this particular case, the item required a very detailed knowledge of encrypting methods; the examiners should consider revising the wording of the item. In general, the test-theoretical analysis cannot determine precise causes of undesired results, but the indicators provide hints to the examiner, which answer alternatives should be reconsidered. In some cases, it may also be appropriate to reflect and change the overall teaching approach for subject areas with low item difficulties or negative discrimination indices.

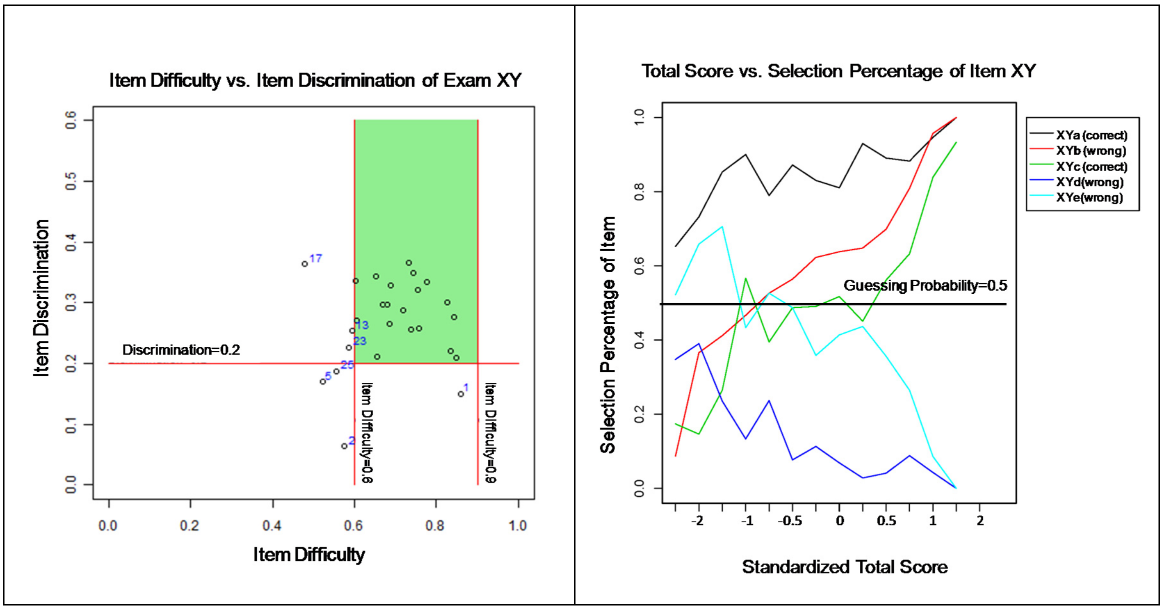

Figure 2: Item analysis graphics

Figure 2 shows two diagrams that visualize test-theoretical indicators. The graphic on the left highlights those items of an exam that fulfil test-theoretical quality criteria (in the coloured background) and those that have potential problems (white background). This diagram might be interesting to give an overview on item quality. The diagram on the right shows whether examinees choose to mark an item as correct depending on their overall performance in the MC test (standardized total score). Thus, the slope of the correct items should go from the lower left to the upper right (e.g. XYb), indicating that more higher performing than lower performing students solve it correctly, while the slope of wrong items should go from the upper left to the lower right of the graphic (e.g. XYe). Item XYa for instance is quite easy for all students, as the slope is not very steep. In addition, graphics on grade distribution and tests of normality of grade distribution could be provided by the system to evaluate whether a test differentiates well between students on all performance levels.

4 Conclusions and further research

The present project was undertaken to design an integrated system that automatically calculates a set of test-theoretical indicators for MC items based on exam results, providing examiners with a feedback on the quality of their MC items. The suggested system design is based on ongoing research efforts to support the process of studying for MC exams within the custom-developed learning management system Learn@WU (Andergassen et al. 2015). At the moment, the system provides the examiners with feedback on item difficulty and discrimination as well as on further details such as multidimensionality, item clustering, or item sensitivity. The provided indicators are not absolute measures per se, they do not brand an item as “good” or “bad” and they do not serve a self-purpose. Even items with indicators out of range can be correctly worded, querying essential knowledge from the course. It is certainly questionable, whether or not it is the purpose of a course and an exam to deliver good discrimination rather than knowledge. However, from our experience the indicators help to direct the examiners’ attention to problematic items in an exam, thus helping them reflect and improve both their MC-writing skill as well as their teaching (Talebi et al. 2013). Several possible directions for future research would be of interest. Opportunities exist for fellow scholars to conduct longitudinal research of item usage as well as to investigate the effect of online-learning behaviour (compare Andergassen et al 2014) on item performance.

5 Bibliography

Andergassen M, Ernst G, Guerra V, Mödritscher F, Moser M, Neumann G, Renner T (2015) The Evolution of E-Learning Platforms from Content to Activity Based Learning. The Case of Learn@WU, In: International Conference on Interactive Collaborative Learning (ICL). Florence, Italy.

Andergassen M, Mödritscher F, Neumann G (2014) Practice and Repetition during Exam Preparation in Blended Learning Courses: Correlations with Learning Results. J Learn Anal 1:48–74.

Bereby-Meyer Y, Meyer J, Flascher OM (2002) Prospect Theory Analysis of Guessing in Multiple Choice Tests. J Behav Decis Mak 15:313–327. doi: 10.1002/bdm.417

Bortz J, Döring N (2006) Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. Springer, Heidelberg, Germany.

Bush M (1999) Alternative marking schemes for on-line multiple choice tests. 7th Annu Conf Teach 3–5.

Chang SH, Lin PC, Lin ZC (2007) Measures of partial knowledge and unexpected responses in multiple-choice tests. Educ Technol Soc 10:95–109.

Cheung D, Bucat R (2002) How can we construct good multiple-choice items? In: Science and Technology Educaiton Conference. Hong Kong.

Dickinson JR (2013) How Many Options do Multiple - Choice Questions Really Have ? 171–175.

Haladyna TM (1992) The Effectiveness of Several Multiple-Choice Formats. Appl. Meas. Educ. 5:73–88.

Kastner M, Stangl B (2011) Multiple choice and constructed response tests: Do test format and scoring matter? Procedia - Soc Behav Sci 12:263–273. doi: 10.1016/j.sbspro.2011.02.035

Kolbitsch J, Ebner M, Nagler W, Scerbakov N (2008) Can Confidence Assessment Enhance Traditional Multiple-Choice Testing? In: Interactive Computer Aided Learning, ICL2008. Villach, Austria.

Lesage E, Valcke M, Sabbe E (2013) Scoring methods for multiple choice assessment in higher education - Is it still a matter of number right scoring or negative marking? Stud Educ Eval 39:188–193. doi: 10.1016/j.stueduc.2013.07.001

Litzenberger M, Punter JF, Gnambs T, et al (2007) Qualitätssicherung bei der Studierendenauswahl mittels lernpsychologisch fundierter Wissensprüfung. Qualitätssicherung und -entwicklung an Hochschulen Methoden und Ergebnisse 23–34.

McCoubrie P (2004) Improving the fairness of multiple-choice questions: a literature review. Med Teach 26:709–712. doi: 10.1080/01421590400013495

Möltner A, Schellberg D, Jünger J (2006) Grundlegende quantitative Analysen medizinischer Prüfungen. Basic quantitative analyses of medical examinations. GMS Zeitschrift für Medizinische Ausbildung 23:Doc53.

Rodriguez MC (2005) Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. Educ Meas Issues Pract 24:3–13. doi: 10.1111/j.1745-3992.2005.00006.x

Sim SM, Rasiah RI (2006) Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. Ann Acad Med Singapore 35:67–71.

Talebi GA, Ghaffari R, Eskandarzadeh E, Oskouei AE (2013) Item Analysis an Effective Tool for Assessing Exam Quality, Designing Appropriate Exam and Determining Weakness in Teaching. 2:69–72. doi: 10.5681/rdme.2013.016

Venter I (2010) Development of a Valid and Reliable Test for Higher-Educated Young Adults Measuring Dietary Fibre Food Source and Health-Disease Association Knowledge. J Fam Ecol Consum Sci /Tydskrif vir Gesinsekologie en Verbruikerswetenskappe 34:10–19. doi: 10.4314/jfecs.v34i1.52906

Wentzel C (2006) A review of INTEGRITY software: An online application to analyze multiple-choice tests and detect test-taking deception. J Sci Educ Technol 15:314–319. doi: 10.1007/s10956-006-9018-2