DO ITEM WRITING BEST PRACTICES IMPROVE MULTIPLE CHOICE QUESTIONS FOR UNIVERSITY STUDENTS?

Author Details:

F. Bertoni, emlyon Business School, Lyon, France

**L.A. Smales, Curtin University, Perth, Australia
University of Western Australia, Perth, Australia

B. Trent CFA Institute, New York, USA

G. Van de Venter, University of Technology Sydney, Sydney, Australia

**Corresponding Author

Abstract:

We conduct a randomized experiment using a series of three finance exams sat by over 1,000 students at an Australian university, to determine the extent to which multiple-choice questions (MCQs) are improved by item writing best practices. We identify deviations from best practices (i.e. "flaws") in MCQs, correct these flaws in a random sample of MCQs and re-administer the exams. Regardless of question difficulty, editing the MCQs according to best practices provides greater clarity for students and increases the proportion of correct responses. We also observe a substantial improvement in MCQ performance related to the effective removal of non-performing distractors. The effect of MCQ editing is larger on those items that students find more difficult. The discriminatory power of MCQs edited according to best practices declines on average, but it increases among students with the lowest scores.

Keywords: Assessment, business education, test validity, multiple-choice questions, best practices

1. INTRODUCTION

Utilized in university education, business school entry (e.g. GMAT) and professional accreditation (e.g. CFA Program), multiple-choice question (MCQ) tests are increasingly ubiquitous. These tests are attractive for a number of reasons. A well-constructed test can be used to test high-order thinking without a dependence on writing skills (Bush, 2006). MCQ tests allow for broad coverage of topics in a single test, accurately discriminate between students of different abilities (Schuwirth and van der Vleuten, 2003), and scores are as reliable as those attained from a constructed-response test (Bacon, 2003). MCQs are particularly attractive in an era of reduced resources and improved technology (Nicol, 2007), and where the automation of marking provides benefits for tests with large enrolment numbers (McCoubrie, 2004; DiBattista and Kurzawa, 2011). For instance, consider the logistical issues arising from marking test papers of the near 200,000 candidates who sat a CFA exam in 2016[1]. Indeed, given the current cost pressures in the education sector it is likely that MCQ testing will become even more prevalent.

However, question writers are often not trained in the process of effective MCQ writing and this raises the likelihood that test takers can misinterpret questions, reducing the validity of the testing process. There are many books, papers, and texts (e.g. Ellsworth et al., 1990; Haladyna, 1997; Haladyna et al., 2002; Kehoe, 1995; Morrison and Free, 2001; Osterlind, 1998) pertaining to best practice design and use of MCQs, but the majority are based on opinion and there is very limited empirical work regarding the effectiveness of the suggested best practices. This paper contributes to rectifying this gap in the literature by empirically testing the impact of best practice designs in MCQs by means of a randomized experiment on a set of actual university exams.

One method of assessing the quality of MCQs is to evaluate the presence of item-writing flaws. Item-writing flaws are violations of accepted item-writing best practices and guidelines, which can readily appear in any of the components of a MCQ[2] and affect student performance. Downing (2002) reports that one third of the questions in a medical exam have flaws and appear to be more difficult than comparable non-flawed items. Downing (2005) and

[1] Source: https://www.cfainstitute.org/programs/cfaprogram/Documents/1963_current_candidate_exam_results.pdf

[2] The conventional format of a MCQ has three components: the stem (question statement), the key (correct answer), and several distractors (incorrect but plausible answers).

Tarrant et al. (2006) find that up to 65% of items in medical exams are flawed, and that this introduces systematic error that reduces exam validity and penalizes some students. Many test banks supplied by textbook publishers are also flawed (Masters et al., 2001) which can lead to problems if they are blindly applied by examiners. Tarrant and Ware (2008) suggest that flawed MCQs are more likely to penalize high-achieving students and not disadvantage borderline students.

Another method to assess MCQ quality is to focus on the characteristics of test responses. DiBattista and Kurzawa (2011) focus on difficulty, discriminatory power, and distractor effectiveness. The difficulty of an item is inversely proportional to the proportion of test-takers who select the correct answer, denoted *p*. The discriminatory power is a measure of the correlation between overall test scores and the score obtained on an individual item. Ebel (1975) suggests this is a major determinant of MCQ quality. DiBattista and Kurzawa (2011) note that items with a low discriminatory power fail to discriminate sufficiently between the high- and low-scoring test-takers and so do not contribute to the overall test quality. They suggest a discrimination coefficient of at least +0.20 is appropriate, and that discriminatory power is greatly impacted by difficulty and the quality of distractors[3]. A good distractor is one that is plausible enough to tempt lower ability test-takers to select it, but will not impact well-prepared students. The literature suggests that at least 5% of test-takers should select a distractor for it to be worthwhile including (Haladyna and Downing, 1993; Tarrant et al., 2009). An effective distractor should also have a discriminatory power that is negatively correlated with overall test scores (i.e., test-takers with high overall scores must select the distractor less often than those with low overall scores). Haladyna and Downing (1993) find that MCQs with more effective distractors are more discriminating, while DiBattista and Kurzawa (2011) find that many of the distractors used in tests at a mid-sized Canadian university function poorly.

This is related to the area of MCQ design that has generated most empirical interest from researchers: the optimal number of distractors. Haladyna et al. (2002) perform a review of the studies published since 1990 and find mixed evidence as to whether reducing the number of options affects difficulty, although there is evidence of better discrimination (Cizek et al., 1998) and increased reliability (Trevisan et al., 1991). More recently, Tarrant and Ware (2010) compare three- and four-option versions of MCQs via a simple strategy (similar to the

[3] Ebel and Frisbie (1991) suggest that items that are very difficult ($p < 0.30$) or very easy ($p > 0.90$) have poor discriminatory power.

one we adopt) of eliminating the distractor with the lowest response rate. They find that three-option items are more highly discriminating items and have fewer non-functioning distractors in addition to the advantage of taking less time to develop. Most convincingly, Rodriguez (2005) performs a meta-analysis over 80-years of research and concludes that three-option items perform best in most educational settings.

Other research into MCQ test quality have identified the importance of the placement of options in logical order (Huntley and Welch, 1993), and the avoidance of "none of the above statements" (Crehan and Haladyna, 1991; DiBattista et al., 2014) although the effect of unfocused stems is limited (Downing et al., 1991).

This paper contributes to the literature by filling two important gaps. First, the research in this field is mostly focused in the area of medical / nursing education, and we can find little trace of empirical work in business education. Our results are based on three finance exams sat by over 1,000 students at an Australian university. Second, we combine the two approaches typically used to measure MCQ test quality: identifying flaws and studying the characteristics of test responses. We adopt a difference-in-difference approach to examine the influence of MCQ writing flaws. We identify the flaws in a set of MCQs used in a baseline sample of students. We then fix the flaws in a systematic manner and re-administer the test. This enables us to perform a series of statistical tests to gauge changes in the characteristics of test responses, and determine the extent to which they are affected by MCQ writing flaws.

Our results demonstrate that the greatest improvement in MCQ performance is related to the effective removal of non-performing distractors. Regardless of question difficulty, editing the MCQs according to best-practices provides greater clarity for students and increases the proportion of correct responses. The effect of MCQ editing is larger on those items that students find more difficult. Intuitively, this make sense because there is little room to improve the proportion of correct responses, or discriminate between high and low achievers, when the vast majority of examinees are already answering the question correctly. The discriminatory power declines on average in the edited MCQs. However, when we look at how the discriminatory power is distributed across the candidates, we observe that editing increases the marginal discriminatory power among the candidates with the lowest scores.

This study is important to academics across various disciplines, publishers of test bank questions and those writing assessments for professional bodies. The paper proceeds as follows. Section 2 describes the data used in this study and outlines our methodology. Section 3 provides our empirical results and accompanying discussion. Section 4 concludes.

2. DATA AND METHODOLOGY

2.1. THE RESEARCH SETTING

The University of Technology Sydney (UTS) is an Australian university formed in 1988 from the earlier NSW Institute of Technology, with its origins tracing back to 1893. This study involves three undergraduate finance subjects, selected to provide a mixture of basic and advanced subjects, taught within the Bachelor of Business degree offered by the UTS Business School.

- **Fundamentals of Business Finance (25300)** is a core first year subject for the Bachelor of Business degree. This subject introduces students to the basic technical and theoretical concepts of finance such as investment and financing decisions. All business students enroll in the subject, irrespective of their subject major, and it is a pre-requisite for subsequent finance subjects.
- **The Financial System (25556)** is the first specialized finance subject for students majoring in finance within the Bachelor of Business degree. The subject covers the payment, financing and market-risk management activities of the financial system.
- **Corporate Financial Analysis (25410)** is a capstone subject for the finance major within the Bachelor of Business degree. It deals with financial statement forecasting and modelling with an emphasis on cash flow reconstructions from financial statements. As a capstone subject, it draws on students' prior knowledge of accounting, economics, finance and quantitative methods.

Each subject is offered twice each year[4] over a standard 11-week semester. The formal assessment structure of each subject comprises of a combination of online quizzes, assignments, group case studies and presentations. Although the assessments and weighting are unique to each subject, all three subjects incorporate a two-hour closed-book final exam at the end of the semester. Each final exam includes of a series of MCQs covering material

[4] In Australia, Semester 1 is typically March – June (Autumn / Fall) and Semester 2 is July – October (Spring). For the purposes of our paper S2 2016 denotes the second semester of 2016 and S1 2017 denotes the first semester of 2017.

delivered during the semester. Summary information for the final exam of each subject is presented in Table 1.

Table 1: Summary of Finance Subjects

Final Examination		**Fundamentals of Business Finance (25300)**	**The Financial System (25556)**	**Corporate Financial Analysis (25410)**
Multiple choice questions		10 questions	14 questions*	16 questions
Weighting of multiple choice component		33.33%	30%	40%
Students taking exam	S2 2016	1,166	409	297
	S1 2017	616	635	304

* Note: The final exams for The Financial System (25556) consisted of 15 questions each for S2 2016 and S1 2017. However, a new topic was introduced into the subject in S1 2017 and the final question in the two multiple choice exams do not correspond. Therefore, for the purposes of our study, this question was dropped from both exams.

2.2. Identification Strategy

Our identification strategy to understand the extent to which item-writing flaws determine the effectiveness of MCQs is based on a randomized experiment combined with difference-in-difference (DID) analysis. DID is a statistical method that can be used to understand the causal effect of a treatment (Lechner, 2010), which in our case is the elimination of item-writing flaws. To the best of our knowledge, this method has not been used extensively in the evaluation of how best-practices affect MCQs, as such it forms an additional contribution of the present paper.

The process is as follows. We select a stratified random sample of approximately half of the MCQs in the S2 2016 exams which forms our set of treated MCQs. From this set, we eliminate any identified item-writing flaw (this step is described in detail in the next section). In the (following) S1 2017 session, for each subject the exam consisted of a combination of our treated MCQs together with the non-treated questions. The DID method then compares how the treated MCQs are affected by the treatment by using the non-treated questions as a control for unobservable factors.

It is useful to describe the DID model more formally (we will follow Lechner, 2010). Let *D* be an indicator variable that identifies questions that are treated, Y be an outcome variable (e.g., the clarity of a question), and X be a set of control variables (e.g., the difficulty of the topic, and the preparation of the student). Let Y^d indicate the hypothetical outcome variable if a given question were treated (Y^1) or not (Y^0). Of course, for a given question either Y^1 or Y^0 is actually observable: if D=1 (i.e., a question is treated) Y^1 is observed; if D=0 (i.e., a question is not treated) Y^0 is observed. However, at least conceptually, we can indicate with: $\theta(x) = E\left[Y^1 - Y^0 \middle| X = x\right]$, the treatment effect conditional on the control variables X. $\theta(x)$ tells us the extent to which the treatment affects a given question after controlling for a number of observable characteristics about the question itself, the exam and the respondent. What is typically estimated by the DID model is the average treatment on treated (ATET) defined as:

$$ATET = E[\theta(x)|X = x, D = 1] \tag{1}$$

The ATET measures the extent to which the outcome variable is affected by the treatment in the treated questions. $\theta(x)$ is not identifiable, because it is a function of variables that are not observable, thus Equation (1) cannot be used directly to estimate the ATET.

Now, let Y_1 and Y_0 indicate the observed values of Y in t=1 (S1 2017) and t=0 (S2 2016) respectively. If the identification assumptions of the DID model are met, $\theta(x)$ can be estimated empirically as follows:

$$\theta(x) = \Big[E\big[Y_1 \;|X = x, D = 1\big] - E\big[Y_0 \;|X = x, D = 1\big]\Big] - \Big[E\big[Y_1 \;|X = x, D = 0\big] - E\big[Y_0 \;|X = x, D = 0\big]\Big] \tag{2}$$

Equation (2) is only based on observable quantities and is then identifiable. We can then simply plug Equation (2) into Equation (1) and calculate the ATET.

The intuition behind Equation (2) is simple: the treatment effect is captured by the change in the outcome variable from S2 2016 to S1 2017 among treated questions (D=1) as compared to the change among non-treated questions (D=0). Non-treated questions are used to create the counterfactual: after controlling for X, the change of the outcome variable in non-treated questions captures how the outcome variable would have changed for treated questions had they not been treated.

Equation (2) requires five identification assumptions (Lechner, 2010), which are discussed in the remainder of this section. The first one is the stable unit treatment value (SUTVA) assumption, which implies that one and only one of the potential outcomes (Y^d) is observable for every member of the population. This assumption is met by design in our study, because every member of the population of respondents either responds to an untreated or to a treated question.

The second assumption is the exogeneity (EXOG) assumption. This assumption implies that the treatment and the control variables X should be exogenous, ruling out the existence of unobserved factors that simultaneously determine the control variables and the treatment. In order to satisfy EXOG, we adopt an experimental setting in which treated units are selected using stratified random sampling. In this way, we eliminate the possibility of endogeneity between the treatment and the controls.

The third assumption is that the treatment had no effect on the pre-treatment population (NEPT). The NEPT assumption for instance rules out cases in which individuals may change their behavior in time $t = 0$ in expectation of a future treatment. In our setting, individual respondents were not aware of the experiment, and their behavior is hence not possibly affected by expectations about a future treatment of some of the questions.

The fourth identification assumption is the common trend (CT) assumption. CT, which is another form of exogeneity, implies that, conditional on X, differences in the expected potential non-treatment outcome over time are unrelated to belonging to the treated or control group. A typical example of how the CT assumption can be violated is self-selection: if treated units self-select into the treatment, the treated group could be different from the non-treated group for unobserved factors that affect the outcome. The result would be a biased and inconsistent estimate of the treatment effect. In our setting, CT is satisfied because, on the one hand, we select treated questions using stratified random sampling and, on the other hand, students cannot self-select into the treated or non-treated exam.

The fifth identification assumption is common support (CS). CS implies that observations with characteristics $X=x$ exist in both the treated and non-treated groups. For all the dimensions of X that are related to student characteristics, CS is automatically satisfied by the fact that, in S1 2017, the same students answer both the treated and the non-treated questions, hence the support for the two groups is exactly the same. For dimensions of X that related to the questions themselves, our experimental setting satisfies CS because we randomly select treated units across strata created from relevant question characteristics, as described in the next section.

2.3. SAMPLE CONSTRUCTION AND MULTIPLE-CHOICE QUESTION QUALITY

Our research methodology follows a process of developing a list of best practice design and use of MCQs, selecting a sample of MCQ from the S2 2016 exams, identifying the flaws in the selected sample, correcting these flaws and finally re-administering the three exams in S1 2017 consisting of sample (treated) and control (untreated) MCQs.

First, we developed a list of best practices for the design and use of MCQs from the literature. The list of 32 best practices (Appendix A) primarily consists of the guidelines presented by Haladyna et al. (2002). Additional guidelines, as well as verification of a number of best practices are contained in Ascalon et al. (2007), Atalmis (2016), Kahn et al. (2013), Malau-Aduli & Zimitat (2012), Morrison & Free (2001), and Rodriguez (2005).

Next, we selected a sample of MCQ to be corrected in accordance to Appendix A. A total of 40 MCQs were available for this study, administered across three exams to 1,872 students in S2 2016. The subsequent S1 2017 exam then consisted of 52.5% treated questions (sample group) and 47.5% untreated questions (control group). We construct the sample using a multi-step process.

Our first step was to classify the MCQs from each of the three S2 2016 exams according to their performance across three criteria that the literature has identified as important:

- **Item Difficulty Index (*p*):** Measures the proportion of students that answered the item correctly[5] and ranges between 0.0 and 1.0. This is sometimes referred to as the "item easiness index". There is some disagreement[6] as to the appropriate range but we follow Mitra et al. (2009) in selecting 0.30 – 0.80 as our acceptable range.
- **Item Discrimination Index (discrimination):** Measures the ability of an item to differentiate between students of higher and lower abilities. Alternatively, a measure of how well an item is able to distinguish between examinees who are knowledgeable and those who are not. Generally, this index ranges between -1 and 1, although its value is normally positive (negative values indicate that high performing students have incorrectly answered the item more frequently than low performing students). We compute discrimination as $DI = [2\,x\,(H - L)\,/\,N]$[7], where H is the number of

[5] There are two methods for this. The first simple divides the number of students who chose the correct answer by the total number of students. The second divides the sum of correct responses from high and low performing groups by the total number of students in the two groups. The proportion of students in each group is typically around 30% and we follow Mitra et al. (2009) and Rasiah & Raja (2006) in using 27%.

[6] For instance, Pande et al. (2013) set an upper-bound of 0.70, Mitra et al. (2009) use 0.80 and DiBattista & Kurzawa (2011) chose 0.90.

[7] An alternative method of computing discrimination, leading to the same numbers is DI = [(H − L)/(N/2)]

examinees in the high-performing group (top 27%) who correctly answer the item, L is the number of examinees in the bottom-performing group (bottom 27%) who correctly answer the item, and N is the total number of students in the high and low groups. We follow DiBattista & Kurzawa (2011) and Pande et al. (2013) in determining that an acceptable level of discrimination is 0.30 or higher.

- **Non-Functioning Distractor:** For a MCQ item to function well, distractors (the incorrect response option) should seem plausible, likely or reasonable to an examinee who is not sufficiently knowledgeable in the content area. We adopt the methodology of DiBattista & Kurzawa (2011) and Gajjar et al. (2014) in determining a distractor as non-functioning if it is chosen by less than 5% of examinees.

Based on the three criteria, and the number of times each MCQ fails to satisfy the criteria, we assign a rating to each MCQ. Table 2 indicates the number of MCQ falling into each category (we provide greater detail in Appendixes B-D).

Table 2: Summary of Rating Categories

Rating	Criteria	Number of questions
A	Satisfies all (3) criteria	19 (47.5%)
B	Satisfies only two (2) criteria	14 (35%)
C	Satisfies only one (1) criterion	4 (10%)
D	Does not satisfy any of the criteria	3 (7.5%)

From the population of 40 MCQs we selected a sample of 21 questions (52.5%) that represent an equal distribution of the four rating categories across all three subjects. Table 3 summarizes this stratified random sample.

Table 3: Summary of the Stratified Random Sample

Rating	Fundamentals of Business Finance (25300)		The Financial System (25556)		Corporate Financial Analysis (25410)		Combined	
	Treated	**Control**	**Treated**	**Control**	**Treated**	**Control**	**Treated**	**Control**
A	1	1	3	4	5	5	**9**	**10**
B	3	3	3	1	3	1	**9**	**5**
C	1	1	0	1	0	1	**1**	**3**
D	0	0	1	1	1	0	**2**	**1**
Total	**5**	**5**	**7**	**8**	**9**	**7**	**21**	**19**
	10		**14**		**16**		**40**	

After our sample of 21 MCQs was selected, we systematically identified the flaws in the selected sample and corrected these flaws in accordance to the guidelines in Appendix A. An example of the changes that were made to a specific question is provided in Appendix B.

The final step was to re-administer the three exams in S1 2017 consisting of the 21 MCQs which were treated and 19 MCQs which were untreated (control).

3. RESULTS

3.1. ITEM QUALITY MEASURES: DESCRIPTIVE ANALYSIS

In this section, we provide a descriptive analysis about the impact of our corrections on the item difficulty (*p*), discrimination power (*discrimination*) and presence of non-functioning distractors (*NFDs*). We are able to analyze our changes at the individual item level in addition to the aggregate exam level.

Table 4 presents some descriptive statistics about how item quality measures changed between the S2 2016 and S1 2017 exam. Students were able to correctly answer more questions in S1 2017, a result that is driven by treated items, for which the p-value has increased by 10 percentage points. The difference in discrimination power is limited and does not seem to be significantly different between treated and non-treated items. The proportion of non-functioning distractors has been reduced for the treated items (-0.48) substantially more than for the control items (-0.05). This results in a substantial increase (+0.24) in the average item's rating category for treated items, whereas we observe a decrease (-0.21) in the rating of control items.

Table 4: Change in item quality measures between S2 2016 and S1 2017

	p	**Discrimination**	**NFD**	**Rating**
All items	0.05	-0.02	-0.28	-0.20
Control items	0.00	-0.02	-0.05	-0.42
Treated items	0.10	-0.03	-0.48	0.00

We then drill down into the individual exams in order to understand the specific drivers of our results, and whether the exam level impacts our results. Table 5 provides results for the Fundamentals of Business Finance exam (25300), the introductory level subject, with our treated items highlighted in bold. Upon the second administration of the exam, the average item became easier and less discriminating, a result primarily owing to the performance of our treated items. The average item rating increased as a consequence of the much lower number of NFDs; the treated (control) group produced six (three) fewer NFDs, contributing to three (two) rating increases from B to A. Two items (one control and one treated) had their rating fall from C to D. In both cases, they had high initial p and the discrimination fell below the cut-off on the second administration.

Table 5: Change in item quality measures between S2 2016 and S1 2017 for the Fundamentals of Business Finance (25300) Exam

25300	*p*	Discrimination	NFD	Rating
Item1	**0.16**	**-0.18**	**-3**	**1**
Item2	-0.08	-0.01	-1	1
Item3	**0.09**	**-0.17**	**-1**	**-1**
Item4	0.03	-0.03	-1	1
Item5	**-0.20**	**0.08**	**-1**	**1**
Item6	0.01	-0.05	-1	-1
Item7	-0.04	-0.02	0	0
Item8	**0.12**	**-0.17**	**-1**	**1**
Item9	**0.17**	**-0.2**	**0**	**0**
Item10	0.08	-0.12	0	0
Average	0.03	-0.09	-0.90	0.30
Average (Control)	0.00	-0.05	-0.60	0.20
Average (Treated)	0.07	-0.13	-1.20	0.40

The results for our intermediate level subject, The Financial System (25556) are shown in Table 6. Again, the average item is more likely to be answered correctly (*p* increased) once treated. In contrast to the introductory level exam, our edits resulted in an average increase in discrimination for the intermediate level subject. In this exam, we had one item that was classified as too easy (very high *p*) and there was no change in the performance of this item. On average, control items dropped by nearly half a rating compared with no change in rating for treated items.

Table 6: Change in item quality measures between S2 2016 and S1 2017 for the Financial System (25556) Exam

	p	Discrimination	NFD	Rating
Item1	0.02	0.04	1	-1
Item2	**-0.04**	**0.09**	**1**	**-1**
Item3	**0**	**0**	**-1**	**0**
Item4	**0.23**	**-0.17**	**0**	**0**
Item5	**0.12**	**0.08**	**0**	**0**
Item6	0.05	0.03	1	-1
Item7	**0.1**	**0.11**	**-1**	**1**
Item8	**-0.01**	**-0.02**	**-1**	**1**
Item9	-0.06	0.04	-1	1
Item10	0.02	0.04	0	0
Item11	0.07	-0.12	0	0
Item12	**0.27**	**0.03**	**0**	**0**
Item13	0	-0.06	0	0
Item14	0.09	0.06	0	0
Average	0.06	0.01	-0.07	-0.21
Average (Control)	0.03	0.00	0.14	-0.43
Average (Treated)	0.10	0.02	-0.29	0.00

Finally, we consider the effect of changes on the exam for the capstone subject, Corporate Financial Analysis (25410) and present the results in Table 7. A familiar story emerges with the number of NFDs declining and the average item becoming easier. However, the performance difference between treated and control items is interesting. In this instance, the control items appear to be slightly less likely to be answered correctly (declining *p*) and less discriminating, while the treated items are more likely to be answered correctly and no change to the discrimination. For this exam, the relative performance of treated items was exactly as we would have expected; higher *p*, similar discrimination, and fewer NFDs. This results in a net rating change of +0.64 for the treated items vis-à-vis the control items.

Table 7: Change in item quality measures between S2 2016 and S1 2017 for the Corporate Financial Analysis (25410) Exam

	p	Discrimination	NFD	Rating
Item1	**-0.01**	**-0.02**	**0**	**0**
Item2	**0.05**	**-0.04**	**-2**	**0**
Item3	**0.01**	**-0.11**	**0**	**0**
Item4	0.01	-0.24	1	-2
Item5	0.07	0.03	0	0
Item6	-0.09	-0.05	0	-1
Item7	**0.03**	**0.02**	**0**	**0**
Item8	**0.17**	**0.2**	**0**	**1**
Item9	**0.23**	**-0.1**	**0**	**-1**
Item10	0.01	-0.19	0	-1
Item11	**0.14**	**-0.07**	**0**	**0**
Item12	-0.08	0.04	0	0
Item13	**0.27**	**-0.07**	**0**	**0**
Item14	0.00	0.08	0	0
Item15	**0.13**	**0.08**	**0**	**0**
Item16	-0.09	0.06	0	0
Average	0.05	-0.02	-0.06	-0.50
Average (Control)	-0.04	-0.03	0.14	-0.86
Average (Treated)	0.12	0.00	-0.22	-0.22

In summary, it appears that our edits have a desirable effect on the performance of students in MCQ exams. The greater clarity of MCQ items results in a greater number of correctly answered questions and reduces the number of non-functioning distractors. The reduction in NFDs is the most effective edit in terms of improving the rating of treated items. The ability to increase the level of discrimination is related to the difficulty of the specific item. For instance, if an item has a *p* below 0.30 then it is typically associated with multiple flaws and thus performance may be improved via careful editing. On the other hand, if the *p* is above 0.80 then there is little room for any discrimination and so our edits do little to improve the rating of such items. This may also be one reason why our treatment had an adverse effect on the discrimination of the introductory level exam (25300) and a more desirable impact on the higher-level exams.

3.2. Treatment Effect On MCQ Difficulty

In the previous section, we have shown that MCQs edited to follow best practice seem to be easier for students. The objective of this section is to verify if this evidence is statistically significant. In order to do that, we use DID analysis as described in section 2.2.

The reason why editing to follow best practice could cause an increase in correct answers is that MCQs that do not follow best practice can be confusing for students, even those who understand the concept that is being tested. Once the item-writing flaws are fixed, and confusion is minimized, more students should be able to answer the question correctly. If best practices actually improve the clarity of a question, we should then expect a positive and significant ATET on the probability that students answer questions correctly.

Therefore, the unit of analysis is the response a student gives to a specific question in an exam. The distribution of the sample across exams and terms is reported in Table 8. Overall, for the three exams over the two semesters we have a total of 3,428 students providing 42,066 individual responses.

Table 8: Distribution of observations

	S2 2016		S1 2017		Both Semesters	
Exam code	**Students**	**Responses**	**Students**	**Responses**	**Students**	**Responses**
25300	1,166	11,660	616	6,160	1,782	17,820
25556	407	5,698	638	8,932	1,045	14,630
25410	297	4,752	304	4,864	601	9,616
All exams	**1,870**	**22,110**	**1558**	**19,956**	**3,428**	**42,066**

Descriptive statistics for the variables used in our DID analysis are reported in Table 9. On average the students correctly answered 59.39% of the questions, as indicated by the dummy variable *Correct Response*. Slightly less than half (47.44%) of the responses refer to S1 2017, as indicated by the variable *S1 2017 dummy*. We indicate treated items with the dummy variable *Treated item*, and set this equal to one for MCQs that were selected for editing. This dummy is the equivalent of D in Equations (1) and (2) and indicates items that were randomly selected to be part of the treatment group. Overall, 51.43% of the responses refer to MCQs in the treatment group. The *Treatment* dummy variable identifies MCQs in the treatment group after they have been edited (i.e., $Treatment = Autumn\,Dummy * Treated\,item$). In order to control for the extent to which a student is prepared (which should be positively associated to the likelihood that, other things equal, the student correctly answers a given question) we calculate *Student grade* as the grade (on a range between 0 and 1) that the student obtained in all MCQs in the exam excluding the focal item. Put differently, *Student grade* is the fraction of correct answers the focal student has in the remainder of the exam.[8]

Table 9: Descriptive statistics

Variable	Obs.	Mean	Std. Dev.
Correct Response	42,066	0.5939	0.4911
S1 2017 dummy	42,066	0.4744	0.4994
Treated item	42,066	0.5143	0.4998
Treatment	42,066	0.2444	0.4298
Student grade	42,066	0.5955	0.2048
Item score in S2 2016	19,956	0.5552	0.1737
Easy item	19,956	0.2186	0.4133
Hard item	19,956	0.2186	0.4133

Note: *Correct Response* is a dichotomous variable equal to one when a given student answers correctly to a given item in a given exam. *S1 2017 dummy* is a dummy equal to one for items in the S1 2017 exam session. *Treated item* is a dummy equal to one for items that were selected for editing. *Treatment* is a dummy equal to one for edited items (i.e., *Treatment = S1 2017 Dummy * Treated item*). *Student grade* is the grade (on a range between 0 and 1) the student obtained in all exam items excluding the focal item. *Item score in S2 2016* is the fraction (between 0 and 1) of students who answered correctly to the item in the S2 2016 term. *Easy item* refers to a value of *Item score in S2 2016* that is below the 27th percentile of items in that exam in the S2 2016 term. *Hard item* refers to a value of *Item score in S2 2016* that is above the 73rd percentile of items in that exam in the S2 2016 term.

Before entering into the complete model to estimate the conditional ATET, it is useful to calculate an unconditional ATET (i.e., the ATET when no X is used as a control variable) with a simple ANOVA table, as reported in Table 10. If we look at MCQs in the control group, we see that students answered correctly to 59.1% of the questions in S2 2016 and to 57.6% in S1 2017. The -1.5% difference is significant at the 5% level. If we look at MCQs in the treatment group, we see that the fraction of correct responses increased from 57.0% in S2 2016 (before the edits) to 63.9% in S1 2017 (after the edits). The 6.9% difference is significant at the 0.1% level. Since the fraction of correct responses increased by 6.9% in the treated group and decreased by -1.5% in the control group, the unconditional DID estimate of the ATET is the difference between these two differences: 8.5% (significant at the 0.1% level). Similarly, we could calculate the ATET by looking at how the difference in the fraction of correct responses between treated and control MCQs changes between the S2 2016 and the S1 2017 session. The difference is -2.2% in S2 2016, but increases to +6.3% in S1

[8] Note that we exclude from the focal MCQ from the calculation of *Student grade*, because it would violate the EXOG assumption of DID by creating an endogenous relationship between the outcome variable and one of the control variable.

2017. Again, the estimated unconditional ATET is the difference between the two differences: 8.5% (significant at the 0.1% level).

Table 10: Unconditional DID analysis

Semester	**Control items**	**Treated items**	**Row difference**
S2 2016	59.1%	57.0%	-2.2%**
S1 2017	57.6%	63.9%	6.3%***
Col. Difference	-1.5%*	6.9%***	8.5%***

Note: The table reports the fraction of correct responses in items that were not selected for editing (*Control items*), and items that were selected for editing (*Treated items*) in the two semesters in which the exam was delivered (*S2 2016* and *S1 2017*). *Col. Difference* is the column difference (i.e., the difference between the means of the *S1 2017* and the *S2 2016* distributions). *Row Difference* is the row difference (i.e., the difference between the means of the *Control items* and the *Treated items* distributions). ***: *p-value*<0.1%; **: *p-value*<1% *: *p-value*<5%.

We can now extend this unconditional analysis by including a number of covariates, as illustrated in Table 11. Because the dependent variable is dichotomous, we report both the results of linear probability models (which are a natural extension of the ANOVA analysis in Table 10, but are not the most appropriate models for binary variables) and probit models (which are more appropriate, but less immediately interpretable, than linear probability models). The results obtained with the two models are entirely consistent in terms of sign and statistical significance.

Table 11: Conditional DID analysis

	OLS			Probit		
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Treatment	0.0846***	0.0834***	0.0868***	0.2193***	0.2368***	0.2609***
	(0.0096)	(0.0093)	(0.0090)	(0.0248)	(0.0266)	(0.0272)
S1 2017 dummy	-0.0218**	0.0048	-0.0192**	-0.0558**	0.0144	-0.0629**
	(0.0069)	(0.0067)	(0.0065)	(0.0177)	(0.0190)	(0.0195)
Treated item	-0.0153*			-0.0391*		
	(0.0066)			(0.0170)		
Student grade			0.5780***			1.7439***
			(0.0107)			(0.0345)
Item fixed-effects	No	Yes	Yes	No	Yes	Yes
N. Obs.	42,066	42,066	42,066	42,066	42,066	42,066

Note: The table illustrates linear probability (columns 1-3) and Probit (columns 4-6) models on *Correct Response*, a dichotomous variable equal to one when a given student answers correctly to a given item in a given exam. *S1 2017 dummy* is a dummy equal to one for items in the S1 2017 exam session. *Treated item* is a dummy equal to one for items that were selected for editing. *Treatment* is a dummy equal to one for edited items (i.e., *Treatment = S1 2017 Dummy * Treated item*). *Student grade* is the grade (on a range between 0 and 1) the student obtained in all exam items excluding the focal item. Item fixed-effects are a set of 39 dummy variables, each identifying one specific item, which are not shown in the table for the sake of clarity. Standard errors are robust and reported in round brackets. ***: *p-value*<0.1%; **: *p-value*<1%.

Column (1) in Table 11 is reported for the sake of clarity, and gives us the same results as in Table 10, estimated using OLS (the ANOVA and OLS approaches are equivalent, see Lechner, 2010). The equivalent baseline model estimated using Probit is reported in Column (4) and gives a perfectly consistent result: the ATET is positive and statistically significant at the 0.1% level. A series of item fixed effects are added in Columns (2) and (5). These variables control for any time-invariant MCQ-specific idiosyncratic characteristic. For instance, MCQs testing different topics may, other things equal, differ in the fraction of correct responses from students (e.g., because some topics are easier than others). Controlling for MCQ-specific characteristics does not affect our main result: the ATET is positive and highly statistically significant (*p-value*<0.1%). Finally, in Columns (3) and (6) we add *Student grade* as an additional control variable. As expected, the variable is positive and statistically significant: students that perform better in the other questions of the exam are, all else equal, more likely to correctly answer any given question on the same exam. Even after controlling for this student-specific characteristic, the estimated ATET remains positive and highly statistically significant (*p-value*<0.1%). To summarize, consistent with our expectations, students correctly answer MCQs that are written following item-writing best-practices more frequently than they do MCQs that do not follow best-practice. The treatment effect is highly statistically significant and large in magnitude.

Finally, we study whether the treatment effect is concentrated in certain item categories. Specifically, we test if the treatment effect is significant in each of the item rating levels and for different levels of item difficulty. We begin by estimating the treatment effect across item rating classes. Based on the rating of items in S2 2016, we create three dummy variables: *A-rated*, *B-rated* and *C/D-rated* (because of the small number of D-rated items, we pool the last two categories). We interact these dummy variables with the *Treatment* variable, to obtain an estimate of the treatment effect across the different rating categories. The results are in Table 12. The first interesting result from this analysis is that, in our most complete models (3 and 6), the treatment effect is positive and significant (*p-value*<0.1%) in each of the three categories. Editing for item best practices makes the items clearer regardless of the initial performance of the item. What is somewhat more surprising is that the treatment effect is larger for items that are rated either A or C/D. The treatment effect is smaller in size and statistical significance (and not statistically significant in models 1 and 4) for items that are in the middle quality category (B). The last two rows of Table 12 illustrate that, in most models, we can reject the null hypothesis that the treatment effect is equal between A- and B-rated items and between B- and C/D-rated items. In other words, clarity seems to be more effective in improving items that were either the best or the worst performing, and less effective in items that are in the middle.

Table 12: Analysis by item rating

	OLS			Probit		
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Treatment * A-rated	0.1190***	0.1707***	0.1765***	0.1739***	0.4374***	0.4798***
	(0.0117)	(0.0130)	(0.0127)	(0.0292)	(0.0349)	(0.0357)
Treatment * B-rated	0.0173	0.0259*	0.0288*	-0.0231	0.0663*	0.0778*
	(0.0120)	(0.0125)	(0.0122)	(0.0290)	(0.0336)	(0.0345)
Treatment * C/D-rated	0.1590***	0.0430***	0.0431***	0.8883***	0.2017***	0.2075***
	(0.0122)	(0.0123)	(0.0122)	(0.0391)	(0.0506)	(0.0521)
S1 2017 dummy	-0.0221**	0.0048	-0.0192**	-0.0558**	0.0144	-0.0631**
	(0.0069)	(0.0067)	(0.0065)	(0.0177)	(0.0190)	(0.0195)
Treated item	-0.0175**			-0.0391*		
	(0.0065)			(0.0170)		
Student grade			0.5787***			1.7496***
			(0.0107)			(0.0345)
Item rating dummies	Yes	No	No	Yes	No	No
Item fixed-effects	No	Yes	Yes	No	Yes	Yes
N. Obs.	42,066	42,066	42,066	42,066	42,066	42,066
(A-rated) – (B-rated)	0.1017***	0.1448***	0.1476***	0.1970***	0.3711***	0.4020***
(B-rated) - (C/D-rated)	-0.1417***	-0.0171	-0.0143	-0.9114***	-0.1355*	-0.1297*

Note: The table illustrates linear probability (columns 1-3) and Probit (columns 4-6) models on *Correct Response*, a dichotomous variable equal to one when a given student answers correctly to a given item in a given exam. *S1 2017 dummy* is a dummy equal to one for items in the S1 2017 exam session. *Treated item* is a dummy equal to one for items that were selected for editing. *Treatment* is a dummy equal to one for edited items (i.e., *Treatment = S1 2017 Dummy * Treated item*). *Student grade* is the grade (on a range between 0 and 1) the student obtained in all exam items excluding the focal item. A-rated, B-rated, and C/D-rated are dummy variables that identify the quality rating of the items in S2 2016. Item fixed-effects are a set of 39 dummy variables, each identifying one specific item, which are not shown in the table for the sake of clarity. Standard errors are robust and reported in round brackets. ***: *p-value*<0.1%; **: *p-value*<1%, *: *p-value*<5%.

Next, we look at differences in the treatment effect across items that differ in difficulty. We identify three MCQ item difficulty categories based on their performance in the S2 2016 exam. We define a variable (*Item score in S2 2016*), which expresses the fraction (between 0 and 1) of students who answered correctly to the MCQ in S2 2016. The we categorize a MCQ item as an *Easy item* if its *Item score in S2 2016* is below the 27th percentile of items in that exam. A *Hard item* is then a MCQ with a *Score in S2 2016* that is above the 73rd percentile of MCQs in that exam in S2 2016. All remaining items are *Medium items*. We interact the treatment effect with the difficulty of the item, and estimate the regression on all 19,956 responses to the S1 2017 exams. The results of estimates using linear probability (Columns 1-2) and Probit (Columns 3-4) models are reported in Table 13.

Table 13: Analysis by item difficulty

	OLS		Probit	
	(1)	**(2)**	**(3)**	**(4)**
Treatment * Easy item	0.0450***	0.0696***	0.2793***	0.3583***
	(0.0107)	(0.0105)	(0.0421)	(0.0427)
Treatment * Medium item	0.0678***	0.0650***	0.1717***	0.1708***
	(0.0082)	(0.0080)	(0.0218)	(0.0223)
Treatment * Hard item	0.1595***	0.1541***	0.4259***	0.4240***
	(0.0124)	(0.0121)	(0.0323)	(0.0330)
Item score in S2 2016	0.9389***	0.8909***	2.5886***	2.5292***
	(0.0236)	(0.0234)	(0.0718)	(0.0727)
Student grade		0.4831***		1.4197***
		(0.0162)		(0.0505)
N. Obs.	19,956	19,956	19,956	19,956
(Easy item) – (Medium item)	-0.0228†	0.0046	0.1076*	0.1875***
(Hard item) – (Medium item)	0.0916***	0.0891***	0.2541***	0.2532***

Note: The table illustrates linear probability (columns 1-2) and Probit (columns 3-4) models on *Correct Response*, which is a dichotomous variable equal to one when a given student answers correctly to a given item in a given exam. *Treatment* is a dummy equal to one for items edited since S2 2016. *Student grade* is the grade (on a range between 0 and 1) the student obtained in all exam items excluding the focal item. *Item score in S2 2016* is the fraction (between 0 and 1) of students who answered correctly to the item in S2 2016. *Easy item* refers to a value of *Item score in S2 2016* that is below the 27th percentile of items in that exam in S2 2016. *Medium item* refers to a value of *Item score in S2 2016* that is between the 27th and the 73rd percentiles of the items in that exam in S2 2016. *Hard item* refers to a value of *Item score in S2 2016* that is above the 73rd percentile of items in that exam in S2 2016. Standard errors are robust and reported in round brackets. ***: *p-value*<0.1%; **: *p-value*<1%; *: *p-value*<5%; †: *p-value*<10%

As expected, *Item score in S2 2016* is positive and highly statistically significant, which indicates a positive correlation between the likelihood that a student correctly answers a question in the S1 2017 semester, and the proportion of students who had correctly answered the question in S2 2016. The estimated treatment is positive and statically significant for each of the three categories of MCQs: regardless of question difficulty, editing the MCQs according to item-writing best-practices provides clarity for students and increases the proportion of correct answers.

The last two rows of Table 13 compare the size of the effect across the different categories of MCQs. There is no strong evidence of a difference in the treatment effect between easy and medium-difficulty MCQs (the results are inconsistent between linear probability models and probit models). However, the results strongly suggest that the treatment effect is larger for the harder items than for the medium-difficulty ones. This result suggests that the benefit of clarity improvement due to item-writing best-practices is particularly significant for those items that are more difficult for the students.

3.3. TREATMENT EFFECT ON THE ITEM CHARACTERISTIC CURVE

In this section we study how editing affects the item characteristic curve (ICC) of MCQs. The ICC is the curve that represents the probability that a student answers correctly to a question as a function of the student knowledge (Lord, 1977), approximated here by the student grade (excluding the focal item). This analysis of the ICC is useful because it allows us to see the extent to which a marginal increase in the Student Grade increases the probability to answer correctly to a given item, which gives us a local measure of the discriminatory power of a question. We estimate both linear probability models and Probit models and include both a specification that is linear in Student Grade and one that includes (Student Grade)^2, which allows us to study the effect on marginal discriminatory power in more detail. The results are reported in Table 14.

Table 14: Analysis of discriminatory power

	OLS		Probit	
	(1)	(2)	(3)	(4)
Treatment	0.1522***	0.0877*	0.4407***	0.2067†
	(0.0184)	(0.0397)	(0.0563)	(0.1220)
S1 2017 dummy	-0.0201**	-0.0180**	-0.0658***	-0.0596**
	(0.0065)	(0.0065)	(0.0196)	(0.0196)
Student Grade	0.6007***	0.1156†	1.8081***	-0.1191
	(0.0119)	(0.0594)	(0.0388)	(0.1857)
Treatment * Student grade	-0.1069***	0.1345	-0.3001***	0.6010
	(0.0253)	(0.1301)	(0.0815)	(0.4132)
(Student Grade)^2		0.4086***		1.6584***
		(0.0474)		(0.1556)
Treatment * (Student Grade)^2		-0.2018†		-0.7747*
		(0.1031)		(0.3414)
Item fixed-effects	Yes	Yes	Yes	Yes
N. Obs.	42,066	42,066	42,066	42,066

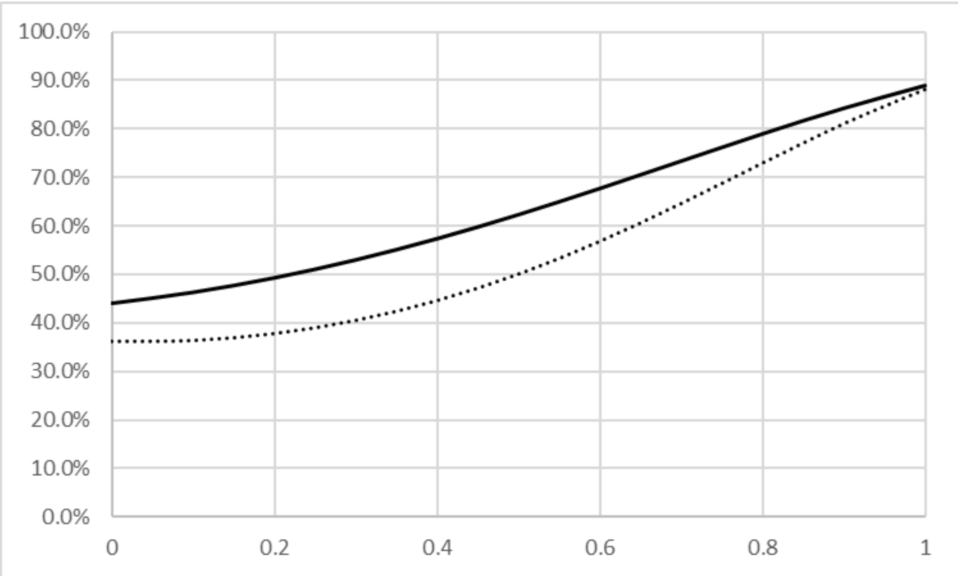
Note: The table illustrates linear probability (columns 1-2) and Probit (columns 3-4) models on *Correct Response*, a dichotomous variable equal to one when a given student answers correctly to a given item in a given exam. *S1 2017 dummy* is a dummy equal to one for items in the S1 2017 exam session. *Treatment* is a dummy equal to one for edited items. *Student grade* is the grade (on a range between 0 and 1) the student obtained in all exam items excluding the focal item. Item fixed-effects are a set of 39 dummy variables, each identifying one specific item, which are not shown in the table for the sake of clarity. Standard errors are robust and reported in round brackets. ***: *p-value*<0.1%; **: *p-value*<1%; *: *p-value*<5%; †: *p-value*<10%.

Models (1) and (3) in Table 14 show that the correlation between the probability of answering correctly to a MCQ and the grade of a student in the remaining MCQs in the exam declines in treated items, as shown by the negative sign (p-value<0.1%) of *Treatment* *

Student Grade. This result confirm the general trend highlighted in Section 3.1: after editing for best-practices, MCQs have a lower discriminatory power. Models (2) and (4) allow us to understand how this average decline in discriminatory power varies across students. The results of these models are better appreciated in graphical form, as illustrated in Figure 1, in which we report, based on Model 4, the ICC and its first derivative (i.e., the marginal increase in this probability with an increase in Student Grade.

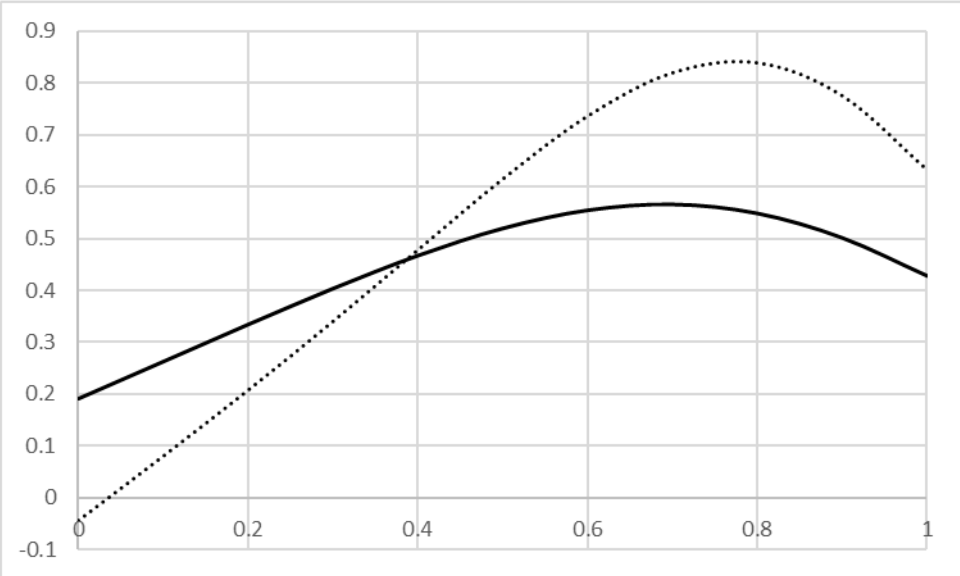
Figure 1: Item Characteristic Curve of treated (solid line) vs. untreated (dashed line) items

Panel A: Item Characteristic Curve



Solid line: treated item; Dashed line: non-treated item

Panel B: Marginal effect of Student Grade on the Item Characteristic Curve



Solid line: treated item; Dashed line: non-treated item

Note: The figure illustrates is based on Model (4) of Table 14 and keeping all other controls at their mean value. Panel A reports the estimated probability to answer correctly. Panel B reports the marginal increase in the estimated probability to answer correctly with a marginal increase in the Student Grade, estimated from Model (4) in Table 14 using the delta method.

Panel A of Figure 1 confirms the finding discussed in the previous sections: students answer correctly more frequently to MCQs that have been edited according to best-practices. The estimated treatment effect, which is the distance between the ICCs is positive across all levels of Student Grade. Panel B of Figure 1 focuses on the marginal effect Student Grade has on the probability of answering correctly to a MCQ. Interestingly, Panel B shows that the greater discriminatory power of non-treated items mostly derives from students with the highest scores, whereas for students with low scores, editing for clarity actually increases discriminatory power. In other words, lack of clarity seems to be particularly confusing for the least prepared students, reducing the ability of MCQs to correctly discriminate among them.

4. CONCLUSIONS

MCQ are widely used as an assessment tool and technological development suggests their usage will continue to grow. It is therefore important to understand the importance of item-writing clarity in assessment outcomes. In this paper, we assess the extent to which following the best practices for item writing results in better MCQs. We do so by building on two streams of literature about MCQs. The first one identifies the best-practices for writing MCQs. The second one determines the main quality measures for MCQs.

This study is among the first to focus on the use of MCQ in business education as we utilize a set of finance exams taken by Australian undergraduate students. A further contribution of our study is that we base our inference on a randomized experiment, which allows us to use difference-in-difference to calculate the treatment effect of writing MCQs according to the best-practices.

Our results demonstrate that editing MCQs according to best-practices has a positive impact on item clarity. Removal of identified MCQ flaws significantly increases the proportion of correct responses. We also find evidence that edited items have fewer non-performing distractors, which increases substantially the performance of MCQs. Editing, however, results in a lower discriminatory power. Further analysis reveals that this effect is driven by a decrease in discriminatory power among the most prepared students, but that the effect of editing on discriminatory power among the least prepared students is actually positive.

This study is important to a range of entities that may use MCQs to assess students, including academics, teachers, professional bodies, and health organizations. Future research may include the application of our methodology to a wider range of test settings. Moreover, the identification strategy used in this paper could be extended to a randomized experiment at the student level, in which each student is randomly assigned a treated/non-treated version of each item. This method would also allow us to have different treated versions of each item, thus testing the relative effectiveness of different categories of edits.

REFERENCES

Ascalon, M.E. et al., 2007. Distractor Similarity and Item-Stem Structure: Effects on Item Difficulty. *Applied Measurement in Education*, 20(2), pp.153–170.

Atalmis, E.H., 2016. Do the Guideline Violations Influence Test Difficulty of High-stake Test?: An Investigation on University Entrance Examination in Turkey. *Journal of Education and Training Studies*, 4(10), pp.1–7.

Bacon, D.R., 2003, Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context, *Journal of Marketing Education*, 25, 31-36.

Bush, M.E., 2006, Quality assurance of multiple-choice tests, *Quality Assurance Education*, 14, 398-404.

Cizek, G.J., K.L. Robinson, and D.M. O'Day, 1998, Non-functioning options: A closer look, *Educational and Psychological Measurement*, 58, 605-611.

Crehan, K.D., and T.M. Haladyna, 1991, The validity of two item-writing rules, *Journal of Experimental Education*, 59, 183-192.

DiBattista, D. and L. Kurzawa, 2011. "Examination of the Quality of Multiple-choice Items on Classroom Tests", *The Canadian Journal for the Scholarship of Teaching and Learning* 2(2), Article 4.

DiBattista, D., J.A. Sinnige-Egger, and G. Fortuna, 2014, The "none of the above" option in multiple-choice testing: An experimental study, *The Journal of Experimental Education*, 82, 168-183.

Downing, S.M., B. Dawson-Saunders, S.M. Case, and R.D. Powell, 1991, The psychometric effects of negative stems, unfocused questions, and heterogeneous options on NBME Part I and Part II item characteristics, Paper presented at the *Annual meeting of the National Council on Measurement in Education*, Chicago.

Downing, S.M., 2002, Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? *Academic Medicine*, 77, S103-S104.

Downing, S.M., 2005, The Effects of Violating Standard Item Writing Principles on Tests and Students: The Consequences of Using Flawed Test Items on Achievement Examinations in Medical Education. *Advances in Health Sciences Education*, 10(2), pp.133–143.

Ebel, R.L., 1975, Can teachers write good true-false items? *Journal of Educational Measurement*, 12, 31-35.

Ebel, R.L., and D.A. Frisbie, 1991, *Essentials of educational measurement*, Englewood Cliffs, NJ: Prentice Hall.

Ellsworth, R.A., P. Dunnell, and O.K. Duell, 1990, Multiple-choice test items: What are textbook authors telling teachers? *The Journal of Educational Research*, 83, 289-293.

Gajjar, S., R. Sharma, P. Kumar and M. Rana, 2014. “Item and Test Analysis to Identify Quality Multiple Choice Questions (MCQs) from an Assessment of Medical Students of Ahmedabad, Gujarat”, *Indian Journal of Community Medicine: Official Publication of Indian Association of Preventive & Social Medicine*, *39*(1), 17–20.

Haladyna, T.H, 1997, *Writing test items to evaluate higher order thinking*, Needham Heights MA: Allyn & Bacon Publishing.

Haladyna, T.H., and S.M. Downing, 1993, How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, 53, 999-1010.

Haladyna, T. M., 2004. *Developing and validating multiple-choice test items.* 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna T.M., S.M. Downing and M.C. Rodriguez, 2002, A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15:3, 309–333.

Huntley, R.M., and C.J. Welch, 1993, Numerical answer options: Logical or random order? Paper presented at annual meeting of *American Educational Research Association*, Atlanta, GA, ED358136

Khan, H.F., K.F. Danish, A.S. Awan and M.A. Anwar, 2013, Identification of technical item flaws leads to improvement of the quality of single best multiple-choice questions. *Pak J Med Sci.* 29(3), 715-718.

Kehoe, J., 1995, Writing multiple-choice test items, *Practical Assessment, Research & Evaluation*, 4(9).

Lechner, M., 2010, The estimation of causal effects by difference-in-difference methods. *Foundations and Trends in Econometrics*, 4(3), pp. 1-224.

Lord, F.M., 1977, Practical applications of item characteristic curve theory, Journal of educational management, 14(2), pp. 117-138.

Malau-Aduli, B.S. & C. Zimitat, 2012, Peer review improves the quality of MCQ examinations. *Assessment & Evaluation in Higher Education*, 37(8), pp.919–931.

Masters, J.C., B.S. Hulsmeyer, M.E. Pike, K. Leichty, M.T. Miller, and A.L. Verst, 2001, Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education, *Journal of Nursing Education*, 40, 25-32.

McCoubrie, P., 2004, Improving the fairness of multiple-choice questions: A literature review, *Medical Teacher*, 26, 709-712.

Mitra, N. K., H.S. Nagaraja, G. Ponnudurai and J.P. Judson, 2009, The levels of difficulty and discrimination indices in type a multiple-choice questions of pre-clinical semester 1, multidisciplinary summative tests. *IeJSME*, *3*(1), 2-7.

Morrison, S. and K. Free, 2001, Writing multiple-choice test items that promote and measure critical thinking. *Journal of Nursing Education* 40: 17-24.

Nicol, D., 2007, E-assessment by design: Using multiple-choice tests to good effect, *Journal of Further and Higher Education*, 31, 53-64.

Osterlind, S.J., 1998, *Constructing test items: Multiple-choice, constructed-response, performance, and other formats*, Boston: Kluwer Academic Publishers

Pande, S.S., S.R. Pande, V.R. Parate, A.P. Nikam and S.H. Agrekar, 2013, Correlation between difficulty & discrimination indices of MCQs in formative exam in Physiology. *South-East Asian Journal of Medical Education*, 7(1), 45-50.

Rasiah, S.S. and R. Isaiah, 2006, "Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper". *Annals Academy of Medicine Singapore*, 35 (2), 67-71.

Rodriguez, M. C., 2005, Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3-13.

Schuwirth, L.W.T., and C.P.M. van der Vleuten, 2003, ABC of learning and teaching in medicine: Written assessment, *BMJ*, 326, 643-645.

Tarrant, M., A. Knierim, S.K. Hayes, and J. Ware, 2006, The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments, *Nurse Education in Practice*, 6, 354-363.

Tarrant, M. and J. Ware, 2008, Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education*, 42(2), pp.198–206.

Tarrant, M., and J. Ware, 2010, A comparison of the psychometric properties of three- and four-option multiple-choice questions in nursing assessments, *Nurse Education Today*, 539-543.

Tarrant, M., J. Ware, and A.M. Mohammed, 2009, An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis, *BMC British Medical Education*, 9, 40.

Trevisan, M.S., G. Sax, and W.B. Michael, 1991, The effects of the number of options per item and student ability on test validity and reliability, *Educational and Psychological Measurement*, 51, 829–837.

APPENDIX A – BEST PRACTICE DESIGN AND USE OF MCQS

Fix	Content Guidelines
1.	AVOID the complex MC (Type K) format
2.	Every item should reflect specific content and a single specific cognitive process
3.	Base each item on important (non-trivial) content
4.	Use novel material to measure understanding and the application of knowledge and skills
5.	Keep the content of each item independent from content of other items on the test
6.	Avoid over-specific or over-general content
7.	Avoid opinion-based items
8.	Avoid trick items
	Style and Format Concerns
9.	Format items vertically instead of horizontally
10.	Edit items for clarity
11.	Edit items for correct grammar, punctuation, capitalization, and spelling
12.	Simplify vocabulary so that reading comprehension does not interfere with testing the content intended
13.	Minimize reading time. Avoid excessive verbiage
14.	Proofread each item
	Writing the Stem
15.	Make directions as clear as possible
16.	Make the stem as brief as possible
17.	Place the main idea of the item in the stem, not in the choices
18.	Word the stem positively, avoid negatives such as NOT or EXCEPT. If negative words are used, use the word cautiously and always ensure that the word appears capitalized and boldface. Phrase choices positively; avoid negatives such as NOT.
	Writing Options
19.	Three option items
20.	Vary the location of the right answer according to the number of options. Assign the position of the right answer randomly.
21.	Place options in logical or numerical order

22.	Keep options independent; choices should not be overlapping
23.	Keep the options homogeneous in content and grammatical structure
24.	Keep the length of options about the same
25.	None of the above should be used sparingly
26.	Avoid using all of the above
27.	Avoid negative words such as not or except
28.	Avoid options that give clues to the right answer
29.	Make distractors plausible
30.	Use typical errors of students when you write distractors
31.	Avoid humor in a high stakes test
32.	Use of peer evaluation of items

APPENDIX B – ANALYSIS OF FLAWS IN EXAM 25300

Original Stats

25300	*p-value*	*p-value* (H&L 27%)	Discrimination	Number of NFD	Rating
(2016)	**(% correct)**				
Item1	0.59	0.60	0.69	3	B
Item2	0.64	0.63	0.64	1	B
Item3	0.82	0.79	0.39	2	C
Item4	0.40	0.45	0.57	1	B
Item5	0.68	0.64	0.59	1	B
Item6	0.86	0.82	0.34	3	C
Item7	0.71	0.67	0.62	0	A
Item8	0.58	0.56	0.66	1	B
Item9	0.47	0.51	0.66	0	A
Item10	0.69	0.65	0.63	1	B
Average	0.64	0.63	0.58	1.30	B

Of the 10 items on the 2016 administration, 2 were rated A, 6 B, and 2 C.

- Eight items had non-functioning distractors.
- All items had strong discrimination.
- Two items were too easy (*p*-value above 0.8). Both of these also had multiple NFDs.

APPENDIX C – ANALYSIS OF FLAWS IN EXAM 25556

Original Stats

25556	*p-value*	*p-value* (H&L 27%)	Discrimination	Number of NFD	Rating
(2016)	**% (correct)**				
Item1	0.79	0.80	0.14	1	C
Item2	0.55	0.54	0.17	0	B
Item3	0.95	0.92	0.12	3	D
Item4	0.38	0.40	0.57	0	A
Item5	0.43	0.46	0.57	0	A
Item6	0.53	0.53	0.66	0	A
Item7	0.35	0.41	0.47	1	B
Item8	0.41	0.47	0.54	1	B
Item9	0.26	0.27	0.28	2	D
Item10	0.43	0.49	0.57	0	A
Item11	0.73	0.67	0.51	0	A
Item12	0.35	0.37	0.41	0	A
Item13	0.44	0.45	0.51	1	B
Item14	0.54	0.56	0.52	0	A
Average	0.51	0.52	0.43	0.64	B 1.86

Of the 14 items on the 2016 administration, 7 were rated A, 4 B, 1 C and 2 D.

- Six items had non-functioning distractors. In three of these cases there were also other problems.
- Four items had discrimination statistics below the 0.3 cutoff. In three of these cases there were also other problems.
- One item was too easy (*p*-value above 0.8). It failed on all other counts as well.
- One item was too hard (*p*-value below 0.3). It failed on all other counts as well.

APPENDIX D – ANALYSIS OF FLAWS IN EXAM 25410

Original Stats

25410	*p-value*	*p-value*	Discrimination	Number of NFD	Rating
(2016)	**(% correct)**	**(H&L 27%)**			
Item1	0.67	0.69	0.46	0	A
Item2	0.82	0.80	0.33	2	C
Item3	0.85	0.83	0.26	1	D
Item4	0.40	0.41	0.53	0	A
Item5	0.61	0.63	0.20	2	C
Item6	0.37	0.40	0.45	0	A
Item7	0.18	0.18	0.00	0	C
Item8	0.40	0.41	0.25	0	B
Item9	0.41	0.41	0.38	0	A
Item10	0.55	0.49	0.46	1	B
Item11	0.50	0.49	0.48	0	A
Item12	0.59	0.59	0.46	0	A
Item13	0.53	0.51	0.41	0	A
Item14	0.32	0.38	0.40	0	A
Item15	0.42	0.44	0.44	0	A
Item16	0.65	0.66	0.38	0	A
Average	0.52	0.52	0.37	0.38	B 1.69

Of the 16 items on the 2016 administration, 10 were rated A, 2 B, 3 C and 1 D.

- Four items had non-functioning distractors. In three of these cases there were also other problems.
- Four items had discrimination statistics below the 0.3 cutoff. In three of these cases there were also other problems.
- Two items were too easy (*p*-value above 0.8). Both of these items failed on one or more other counts as well.
- One item was too hard (*p*-value below 0.3). It also failed on the basis of discrimination.

APPENDIX E – EXAMPLE OF APPLYING BEST PRACTICES

The following example of changes that were made to a specific question illustrates the process and subsequent results.

The following question appeared on the S2 2016 exam for the subject The Financial System (25556):

Our study of spot and forward interest rates reveals:

(a) *implicit forward rates are embedded in the second period of two-period rates;* [9]
(b) *spot interest rates commence two days after their contracts are traded;*
(c) *an inverse yield curve results when* ${}_1r_2 > {}_0r_1$*;*
(d) *both (b) and (c) only are correct.*

A number of best practice violations were identified, which includes the following:

- Every item should reflect specific content and a single specific cognitive process (Fix no. 2, Appendix A)
- Three option items (Fix no. 19, Appendix A)
- Keep options independent; choices should not be overlapping (Fix no. 22, Appendix A)
- Keep the options homogeneous in content and grammatical structure (Fix no. 23, Appendix A)

The MCQ was subsequently changed and appeared on the S1 2017 exam as follows:

Spot interest rates most likely:

(a) *embed implicit forward rates.*
(b) *commence two days after their contracts are traded.*
(c) *are expected to increase if the yield curve is inverted.*

The following illustrates the impact of the changes on the performance of the question:

[9] The correct answer (key) is A in both versions of the question.

Semester	***p-value***	**Discrimination**	**Non-functioning Distractors**	**Rating**
S2 2016	0.41	0.47	1	B
S1 2017	0.47	0.58	0	A