COULD OPEN-ENDED QUESTIONS BE REPLACED BY MULTIPLE-CHOICE QUESTIONS IN STATISTICS EXAMS?

Evert Jan Bakker and Hilde Tobi
Biometris, P.O. Box 16, 6700 AA Wageningen, The Netherlands
evert-jan.bakker@wur.nl

Exams of a secondary service course on Statistics at University, covering t-tests, linear regression, 1- and 2-way ANOVA, and non-parametric tests, contained two parts: one with open-ended questions and one with multiple-choice questions, each with 45 credits max. We have student scores (per item) of approximately 4000 first attempts for 12 different exams, some of which are used several (≤ 4) times. Main question is whether we could use multiple-choice questions only in our exams. Besides showing some descriptive statistics for the two parts, and analyses of the relationship between scores for the two parts, we will present results of a series of factor analyses to identify whether the items in the two sets represent different latent traits.

INTRODUCTION

In a secondary service course on Statistics at Wageningen University, the Netherlands, we cover t-tests, linear regression, 1- and 2-way ANOVA, ANCOVA, and non-parametric tests. The learning objectives fall under the headings of "application" and "understanding". Our 3-hour exams used to have open-ended questions only. With increasing student numbers we wondered if we might as well use multiple-choice (M-C) questions. Since the end of 2013 the exams contain

The open-ended part of the exam typically consists of 3 problems with a number of sub-questions (a, b1, b2, etc., some 18 in total). For each of these items, the maximum score (1 – 4) is communicated to the students in the exam. The maximum scores add up to 45. The problems are in the form of a study for which the data are analysed using SPSS and computer output is given for 1-way or 2-way ANOVA, linear regression, or ANCOVA. In each exam, students are asked to describe the design of a given experiment once (e.g. mention treatment factors, experimental units, blocks, covariates), to interpret computer output, carry out tests, and draw conclusions on factor effects. They are to carry out F-tests and t-tests, discuss interaction, do pair-wise comparisons between factor levels, formulate conclusions, calculate R^2, R^2_{adj}, or a standard error, and answer questions to test understanding of the conclusions, and methods used.

The multiple-choice part (25 items, each with 4 answer options) contains one ANOVA, ANCOVA or regression problem, with 8 or 9 items. These are comparable to the open-ended questions in learning goals and content they cover. Secondly, there are some 8 items on proportions (estimate one proportion, or the difference between two; binomial test, Fisher exact test) and other count-data suitable for the application of chi-square tests. Often these items come in the context of one problem (with e.g. a 3x4 table of counts), where the elements above are addressed by looking at all data, at a subset of the data, or at a new table formed after merging some of the rows or columns. Thirdly, two items with sample size calculations are included. Finally, a few items cover elements that should be addressed in the exam, but, so far, were not, or according to inspiration

Note that the MC-score = (NCA - 6) *2.5, with NCA = Number of Correct Answers. MC-score has a maximum of 45, and minimum 0; 6 is subtracted as correction for 'guessing'. OpenScore is the score for the open-ended items. TotalScore = MC-score + OpenScore, and the grade for students is TotalScore /10 +1, ranging from 1 to 10. A TotalScore of 45 or more is a "pass" for the students.

DATA FROM THE EXAMS

We obtained student scores (per item) in over 20 exams. Some of these exams are (partial) copies of previous exams. In most exams there are two versions (A and B), that are different in the order of the M-C questions only. In the data set, the scores for the M-C questions of version B are placed in the order of the corresponding A-questions, so that the scores in one column are scores for the same question. As examples, the exams 1610 (October 2016), 1612, 1710 and 1712 are used in the next analyses, as they are recent and have the largest student numbers.

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

COMPARING MC-SCORE AND OPEN- ENDED SCORE IN EXAMPLE EXAMS

On a total maximum score of 45+45=90, the average scores for the open-ended questions (OpenScore) are usually close to 26 (Table 1), where the average scores for the M-C questions (MCscore) were somewhat lower: between 19.8 and 22.6. The September 2017 exam gave an average score of 28.6 and 22.6, respectively.

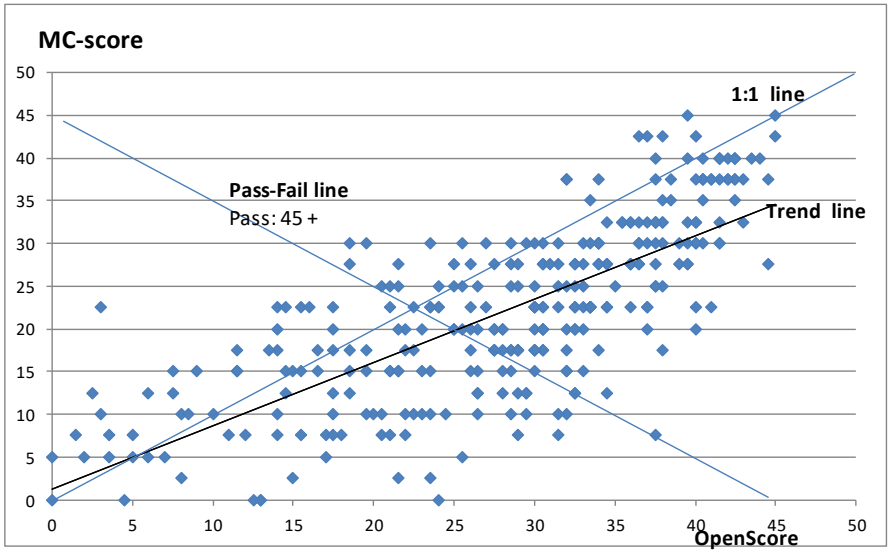

Figure 1. MC-score vs OpenScore for exam 1710 with Pass-Fail line, 1:1-line and trend line.

We drew a graph like the one in Figure 1 for each exam. Based on these we observed that students with lower marks tend to score "extra low" on multiple-choice questions.
The students we "worry" most about are those who have a relatively high score for the open-ended questions, say ≥ 28, and yet fail the exam. Such a phenomenon was never observed in the opposite direction, e.g. the maximum MC-score for failing students in exam 1710 is 22.5.
Using Figure 2, the MCscore and Open-Score relationship is analysed further.
LOWESS curve fit-tings, (locally weighed Scatterplot Smoother, or local polynomial regression, a tool in the SPSS graphics mode) indicate that the slope of the relationship increases with increasing OpenScore. We estimated a model for subsets of students with OpenScore < 20, and with OpenScore > 25.

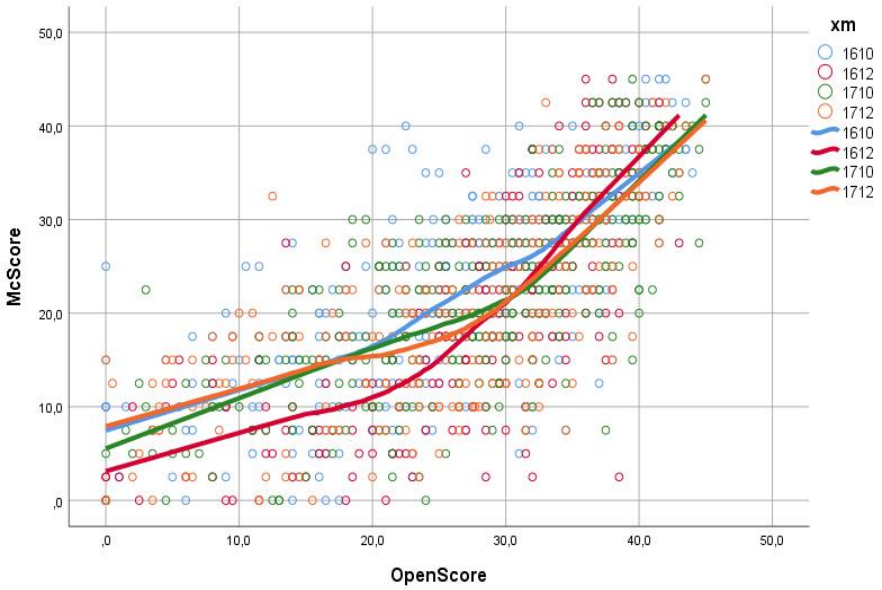

Figure 2. Relationships MC-score vs OpenScore for four exams

We did this twice, once in a one-slope model, once in a model that allowed different slopes per exam. Indeed, the slope of MC-score vs OpenScore is higher for the subset with high OpenScores than for the subset with low OpenScores (Table 1).

Exam	*Mean score (sd)*		*TS vs MCS*		*Slope MCS vs OS*		*Correlations*		
	OS	MCS	Slope	Intercept	OS<20	OS>25		OS	MCS
1610	25.8	22.0	1.68	10.74	0.28	0.97	OS		0.7
(n=352)	(9.9)	(10.1)					TS	0.92	0.92
1612	26.9	19.8	1.62	14.66	0.35	1.36	OS		0.73
(n=224)	(9.9)	(11.8)					TS	0.92	0.94
1710	28.6	22.6	1.75	11.76	0.6	1.18	OS		0.74
(n=356)	(10.2)	(10.2)					TS	0.93	0.93
1712	26.2	20.7	1.73	11.21	0.53	1.21	OS		0.69
(n=247)	(10.7)	(10.1)					TS	0.92	0.92

Table 1. For four example exams: Means of OpenScores (OS) and Multiple ChoiceScores (MCS) Regression coefficients TotalScore (TS) vs MCS, slopes MCS vs OS for two segments of the data set, and correlations between OS, MCS, and TS

Of particular interest are the correlations between the scores for the two exam parts and the overall score, shown in Table 1. The correlations of each of the parts with total score are all above 0.9. To conclude that the MCscore is a good predictor for TotalScore, the MCscore and TotalScore needs to be constant across exams (Table 1): in three of the four exams the relationship is very similar.

FACTOR ANALYSIS

For the factor analyses (PCA method) we combined data from corresponding exams regardless of item order. Student scores were transformed into a percentage of the maximum score (0-100), because occasionally the maximum scores had changed between exam versions. MCscores are either 0 or 1. For the factor analyses, we created 5 data sets. In 50 cases students made the same exam twice. We maintained the student scores of the last exam only in the data set. See Table 2 for details on the 5 data sets.

The exam versions within one data set were not always exactly the same: sometimes an item was removed from the previous version of the exam. We included in our analyses only items that were answered by all students, so the numbers of students given in Table 2 are also the numbers of observations used in the analyses. We did a PCA for open-ended items only, for M-C items only, and for all items together.

Data set	1	2	3	4	5
Exams (yymm)	1712 1604 1404 1408	1710 1608 1504	1612 1602 1506	1610 1508	1412 1406 1402
N	526	525	381	395	382
Nopen	19	17	18	16	17
Nmc	25	25	24	23	25

Table 2. Corresponding exams in each of 5 data sets. N= number of students per data set, Nopen = number of distinct open-ended items per data set, and Nmc the number of included multiple-choice questions.

Results of Factor Analysis

Table 3 presents results of all PCA analyses. Across exams and item sets, the number of components identified based on the Kaiser-criterion (Eigenvalue > 1), and % of variance explained

were similar. For the open-ended items between 2 and 4 components could be distinguished, for the multiple-choice items 8 or 9. When items were combined the sum was more or less found.
The percentages of total variation explained by the first component ranged from 35-42% for the open-ended questions to a mere 12-16% for the multiple-choice ones. In the combined analysis, 20-23% of variance was explained by the first component (not given in the Table 3). All but four open-ended item scores had loadings > 0.3 on the first component, while between 30% and 70% of the mc-items had loadings < 0.3.
The full component solution explained a bit over 50% of overall scores variability. Assuming that overall scores are the best available estimate, it appears that the open-ended items better represent the underlying latent trait "mastering the course material" or another general latent trait. This is further investigated by the correlations of first components with overall exam score. The first component (PC1) in the combined analysis correlated between .994 and .998 with overall exam score; if PC1 is only based on open-ended items, the correlation is between .96 and .98, but for mc-questions it is somewhat lower, namely between .83 and .89. When looking at internal reliability analysis, Cronbach's alpha values ranged from .85 to .90 for open-ended items, between .68 and .77 for multiple-choice items, indicating higher internal validity for the scores from open-ended questions. For all items combined, the range was .83 to .88 (not in Table 3).

	Open-ended items					Multiple-choice items				
Data set	NC	%P1	%T	r(S,P1)	Cr α	NC	%P1	%T	r(S,P1)	Cr α
1412	2	41	47	0.98	0.9	8	12	50	0.83	0.68
1610	3	35	49	0.96	0.87	8	15	51	0.84	..73
1612	3	31	49	0.96	0.85	8	16	51	0.88	0.77
1710	2	42	49	0.98	0.91	8	15	48	0.89	0.74
1712	4	39	58	0.97	0.91	9	14	51	0.86	0.73

Table 3. Factor analysis results for analyses applied to 5 data sets. NC=Number of components extracted (criterion: EV<1). %P1=% of total variation explained by PC 1. %T = % of total variation explained by all components; r(S, P1) = correlation of total score S and PC1; Cr α = Crombach's α.

GENERAL DISCUSSION

With increasing student numbers we wondered if we might replace the open ended questions in the exams by multiple-choice questions. The results presented suggest that the correlation between the totals of the M-C scores and the open-ended question scores is sufficiently high to assume that they (also) measure a common trait. In general, the principal component analyses supported the presence of one common trait despite the range of topics covered in this course. However, it also showed that a substantial portion of the item variation is not explained by this one component.
The variance not explained by the first component does not necessarily reflect noise. As particularly open-ended questions loaded on the first component, the first component may reflect general statistical understanding whereas next components may be more related to the capability to carefully read and to understand specific statistical terminology as required for the M-C questions.

CONCLUSION

Concluding, the use of only multiple-choice questions may be defended on the basis of the results, but the open question scores do, in our exams, a better job of predicting the total score, and are more internally reliable. So, at present, a price would have to be paid for changing to multiple-choice questions only. It is, however, possible, that not the type of question as such is the cause of the difference, but that the overall quality of the multiple-choice questions is not as high yet as that of the open-ended question. As for us, we are likely to continue with the mixed answer format exams.