Assessing engineering students with multiple-choice exams
theoretical and empirical analysis of scoring methods

T. De Laet[1], **J. Vanderoost**[1], **R. Callens**[1], **and R. Janssen**[2]
[1]Faculty of Engineering Science, Tutorial Services &
Leuven Engineering and Science Education Centre (LESEC)
[2] Faculty of Psychology and Educational Sciences
KU Leuven
Leuven, Belgium
E-mail: Tinne.DeLaet@KU Leuven.be

Conference Key Areas: Engineering Education Research, Gender and diversity, “I feel brilliant”

Keywords: assessment, multiple-choice questions, risk-aversion

INTRODUCTION

Multiple-choice questions (MPQ) are considered an objective way for testing large groups, and allow for fast feedback. A drawback of MPQ is that students can gain marks by guessing, and that depending on the marking method personality traits such as risk aversion might influence the total score. A variety of marking methods for MPQ are available [1]–[4] each with their own advantages and disadvantages. Methods such as negative marking try to discourage students from gaining marks by guessing by introducing a penalty for a wrong answer. Other methods such as standard setting correct for guessing by increasing the threshold for passing. Methods such as elimination marking, allow rewarding partial knowledge.
This study tackles one important concern: the fair and objective marking of multiple-choice exams with a special focus to the influence of “guessing”. The focus is on comparing three widely used marking methods for MPQ: negative marking (NM), standard setting (SS), and elimination marking with adapted score rule (EMA) [5]. This paper is the first one to present results both theoretical and experimental related to EMA. This paper uses a theoretical framework that combines statistics, economic and psychometric methods to study the effect of risk-aversion and ability for NM, SS, and EMA. Secondly, it presents the result of an empirical analysis of two exams of first year engineering students with EMA, and the results of a survey where students compare NM and EMA.

1 RELATED WORK

Literature, both from the educational and psychological field of assessment as from the economic research field of decision making under uncertainty, has however indicated that different marking methods can have a different influence on students’ response behaviour and the obtained score depending on their personality traits. In particular it has been shown that negative marking disadvantages risk-averse

students [1], [4], [6]. As some studies indicate that female students are more risk-averse, marking methods can introduce an unwanted gender bias [7], [8] as female students tend to leave more questions blank when confronted with a punishment for wrong answers.
Bond et al. [9] showed that elimination marking, a method that rewards partial knowledge but still introduces a penalty for guessing, does not introduce a gender bias in life sciences. Moreover, they found that this method increases student performance and satisfaction and reduces anxiety.
Some previous studies [5], [6], [10] used prospect theory to analyse guessing in multiple-choice tests. [6] is the first to show the usefulness of prospect theory in the non-financial context of multiple-choice exams. They showed that the behaviour on multiple-choice exams can be predicted by prospect theory. Additionally, [5] predicted the influence on the expected exam score for NM and EMA as a function of risk-aversion. In this paper, we take a step further by combining prospect theory with psychometry to analyse scoring methods depending on risk-aversion ánd ability.

2 SCORING METHODS

This paragraph explains the different scoring methods that are subject of this paper. *Table **1*** show the different answering patters, respectively, that the different scoring methods can handle, and they use an example with four alternatives ($\mathrm{n} = 4$) as a clarification. n is the number of alternatives, N is the number of questions. In the analysis it is assumed that each question has that only one and exactly one alternative is the correct answer.

2.1 Negative marking

In NM, the students can either indicate the alternative they to be correct or leave the question blank. A wrong answer receives a penalty $\frac{-1}{n-1}$, a blank answer is scored neutrally (0). The scoring rule for negative marking is:

$$score = \frac{score_{max}}{N}\left(y - \frac{1}{n-1}(N - y - b)\right)\cdot, \tag{1}$$

where y is the number of correct answers (the raw score) and b the number of blank answers. $(\mathrm{N} - \mathrm{y} - \mathrm{b})$ is therefore the number of wrong answers. For negative marking the threshold for passing is typically $c = \frac{N}{2}$, i.e. half of the questions.

2.2 Standard setting

In SS, the students have to indicate the alternative they believe to be correct. To account for guessing SS is often combined with a higher threshold for passing:

$$c = N\ \frac{n+1}{2n}. \tag{2}$$

This threshold is equal to the expected number of correct answers of a student that knows half of the questions $\frac{N}{2}$.The basic scoring rule for SS is (with y the number of correct answers and $score_{max}$ the maximum score as student can obtain, e.g. 10):

$$score = \frac{score_{max}}{N}\ y. \tag{3}$$

Additionally, the final score can be corrected for guessing as:

$$score = score_{max} * \left(1 + \frac{1}{N - c}(y - c)\right). \tag{4}$$

2.3 Elimination marking

In elimination marking the students have to eliminate the alternatives they believe to be incorrect [11]. Arnold and Arnold [12] introduced a scoring procedure for elimination that allows for partial knowledge and also allows the examiner to control the expected gain due to guessing. The EMA presented in this paper is a special case of the proposed scoring procedure, with a "fair penalty", i.e. a penalty resulting in an expected gain of zero when guessing [12]. EMA allows the student to indicate doubt (partial knowledge) by eliminating fewer than $n-1$ alternatives. If the student does not indicate doubt (and eliminates all but one alternative), the scoring is exactly the same as negative marking. The scoring rule for EMA is:

$$score = \frac{score_{max}}{N}\left[\sum_{x=0}^{n}\left(N_x \frac{x}{(n-x)(n-1)}\right) - \frac{1}{n-1}(N-y-b)\right], \quad (5)$$

with y the number of non-blank answers where the correct alternative is not eliminated (no misconcept), x the number of alternatives eliminated, and N_x the number of answers where x distractors are eliminated (and the correct alternative not).

Table 1: Answering patterns for negative marking (NM), standard setting (SS), and elimination marking with adapted scoring rule (EMA) and the possible scores for the answering pattern.

			possible scores	
	possible answering patterns	**[A B C D]**	**correct answer indicated/ not eliminated (e.g. A= correct)**	**correct answer not indicated/eliminated (e.g. D=correct)**
NM	no doubt	[1 0 0 0]	1 (full knowledge)	$-1/3$ (misconcept)
	blank	[0 0 0 0]	0 (no knowledge)	0 (no knowledge)
SS	no doubt	[1 0 0 0]	1 (full knowledge)	0 (misconcept or no knowledge)
EMA	no doubt	[0 1 1 1]	1 (full knowledge)	$-1/3$ *(partial misconcept 1)*
	doubt two alternatives	[0 0 1 1]	$1/3$ (partial knowledge 1)	$-1/3$ (partial misconcept 2)
	doubt three alternatives	[0 0 0 1]	$1/9$ (partial knowledge 2)	$-1/3$ (total misconcept)
	blank	[0 0 0 0]	0 (no knowledge)	0 (no knowledge)

3 METHODOLOGY & FINDINGS

This paper takes a theoretical and an empirical approach.

3.1 Theoretical analysis

The theoretical analysis consists of three parts: an analysis of SS, the reliability of the scoring methods, and the influence of ability and risk-aversion.

3.1.1 Standard setting with adapted scoring

By substituting the adapted threshold c (2) into the scoring rule (4) and doing simple mathematical manipulations the rewritten scoring rule is obtained:

$$score = \frac{score_{max}}{N}\left(y - \frac{1}{n-1}(N-y)\right). \quad (6)$$

From this formula it becomes clear that the $N-y$ non-correct answers (these are the wrong answers and the blank answers) are punished with a "correction" of $\frac{-1}{n-1}$. When comparing this with the scoring rule or negative marking (1) it is clear that in SS all non-correct answers receive the same punishment as the wrong answers in NM.

Therefore, from a scoring point-of-view SS with adapted scoring is equivalent to NM without the possibility to leave questions blank (or to consider blank answers as wrong answers).

3.1.2 Reliability of the scoring method

The reliability of the scoring method is affected by the additional variance of the scores apart from the variance due to individual differences in ability and risk-aversion. If students do not know the answer, they may "guess". This guessing introduces a variance on the scores of students. Both NM and EMA allow leaving questions blank, such that students can avoid entering the "game of gambling": they can choose for a certain score of 0 by leaving the question blank. On the other hand, in SS blank answers receive the same score as wrong answers. Consequently, students are advised to choose an alternative for every question and, hence, students will enter the "game of gambling" to guess the correct answer. This introduces a variation on the scores of students with the same characteristics, decreasing the reliability of SS.

To study the additional error variance, the following scenario is investigated. Different students with the same characteristics know the answer of x out of N questions, and it is assumed that these answer to this x questions are correct (no misconcept). For the remaining $N - x$ questions, it is assumed that the students guess randomly (no partial knowledge). The score corresponding to getting x and additionally a of the $N - x$ guessed answers right is (without adapted score according to (4)):

$$score(a,x) = \frac{score_{max}}{N}\,(x+a). \tag{7}$$

The probability of getting a of the $N - x$ answers right (and therefore of obtaining $score(a,x)$) is the binomial distribution (when assumed that this probability is not influenced by ability):

$$p(a|x) = p(score(a,x)|x) = \binom{a}{N-x}\left(\frac{1}{n}\right)^{a}\left(\frac{n-1}{n}\right)^{(N-x)-a}. \tag{8}$$

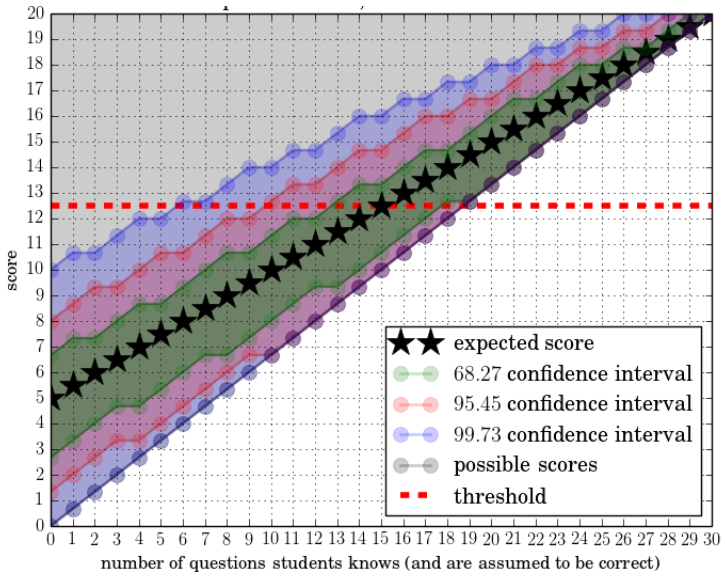

Fig. 1: Theoretical analysis of expected score and variance on expected score for standard setting (SS) with final score according to (4), 30 questions, 4 alternatives

From this binomial distribution the confidence bounds can be calculated, resulting in an indication on the expected variance on the scores of students with the same

ability x. *Fig. 1* graphically presents the results for $N = 30$ questions and $n = 4$ alternatives. This analysis shows that for a typical number of questions of multiple-choice exams in engineering (20-60), the expected error variance on the score when using SS is higher than NM, which offers a blank possibility. This reduces the reliability of the SS for scoring multiple-choice exams.

3.1.3 Ability and risk-aversion

The third part of the theoretical analysis uses statistics, probability theory, prospect theory [13], [14], and the partial credit model [15], [16] to simulate a student's answer on a MPQ for different scoring methods, depending on his/her risk-aversion and ability.

A two-step approach is used.

Step 1 (ability → probability knowledge levels) gives as a function of the ability, the knowledge levels a student is expected to show for a particular question. The model is inspired by the partial credit model [16]. *Table 2* defines the different knowledge levels for $n = 4$. *Fig. 2* shows $p(knowledge\ level)$ for the question with parameters studied in this paper.

Table 2: Different knowledge levels

knowledge level	description
perfect knowledge (FK)	knows correct answer
partial knowledge type 1 (PK1)	doubt between correct answer and distractor
partial knowledge type 2 (PK2)	doubt between correct answer and two distractors
no knowledge (NK)	doubt between all alternatives
partial misconcept type 1 (PM1)	thinks one distractor is correct answer
partial misconcept type 2 (PM2)	doubt between two distractors
total misconcept (TM)	doubt between three distractors

Table 3: Knowledge types and corresponding expected value of different answer patterns for EMA for n=4 given the knowledge level, i.e. $\boldsymbol{E(v_{AP}|KL)}$. $\boldsymbol{v_{KL}}$ is short for $\boldsymbol{v(score_{KL})}$, with $\boldsymbol{score_{KL}}$ *obtained from Table 1*.

knowledge level (KL)	answer patters (AP): no doubt	doubt two alternatives	doubt three alternatives	blank
FK	v_{FK}	v_{PK1}	v_{PK2}	v_{NK}
PK1	$\frac{1}{2}v_{FK} + \frac{1}{2}v_{PM1}$	v_{PK1}	v_{PK2}	v_{NK}
PK2	$\frac{1}{3}v_{FK} + \frac{2}{3}v_{PM1}$	$\frac{1}{3}v_{PM2} + \frac{2}{3}v_{PK1}$	v_{PK2}	v_{NK}
NK	$\frac{1}{4}v_{FK} + \frac{3}{4}v_{PM1}$	$\frac{1}{2}v_{PM2} + \frac{1}{2}v_{PK1}$	$\frac{1}{3}v_{TM} + \frac{3}{4}v_{PK2}$	v_{NK}
PM1	v_{PM1}	$\frac{2}{3}v_{PM2} + \frac{1}{3}v_{PK1}$	$\frac{1}{4}v_{TM} + \frac{2}{3}v_{PK2}$	v_{NK}
PM2	v_{PM1}	v_{PM2}	$\frac{1}{2}v_{TM} + \frac{1}{2}v_{PK2}$	v_{NK}
TM	v_{PM1}	v_{PM2}	v_{TM}	v_{NK}

Step 2 (probability knowledge levels, risk-aversion → answer pattern) gives as a function of the risk-aversion and the knowledge levels obtained in step 1, the expected answer on the question and the associated expected score and variance on the score. To this end, prospect theory is used. Prospect theory is a behavioural economic theory that describes the way people choose between probabilistic

alternatives that involve risk, where the probabilities of outcomes are known. The theory states that people make decisions based on the potential value of losses and gains rather than on the final outcome (here: the score), and that people evaluate these losses and gains using certain heuristics. Depending on individual characteristics people attach a personal "value" $v(x_i)$ to an outcome x_i. In case of a multiple-choice question the outcome x_i is the scaled score on the question. When making a decision under uncertainty, people take into account the probabilities $p(x_i)$ they attach to the different outcomes x_i. In this model it assumed that people attach a different value to negatives outcomes, depending on their risk-aversion λ. In the applied model the possible outcomes are the scores of the different answer patterns (*Table 1*). In order to apply prospect theory, given the different knowledge levels the expected values of the different answer patterns have to be calculated. As an example *Table 3* shows the formulas for the expected values for EMA. The expected value of a particular answer pattern (AP) is then:

$$E(v_{AP}) = \sum_{KL} p(KL) \,.\, E(value_{AP}|KL). \qquad (9)$$

Fig. 3 shows an example of step 2 where for a particular ability the expected value of the different answer patterns is shown in function of risk aversion for EMA.

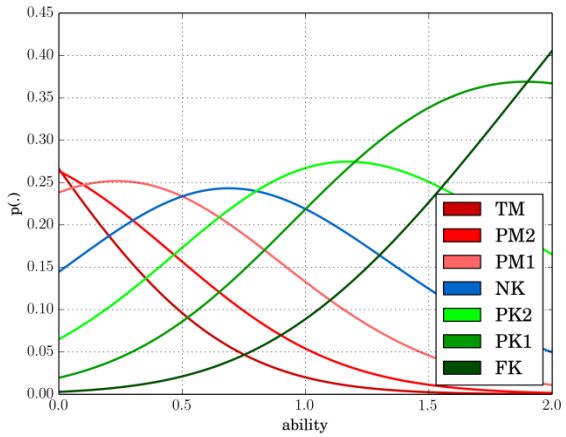

Fig. 2: Model relating the ability of a student to probabilities of different knowledge levels shown by the student for a particular MCQ (step 1), 4 alternatives.

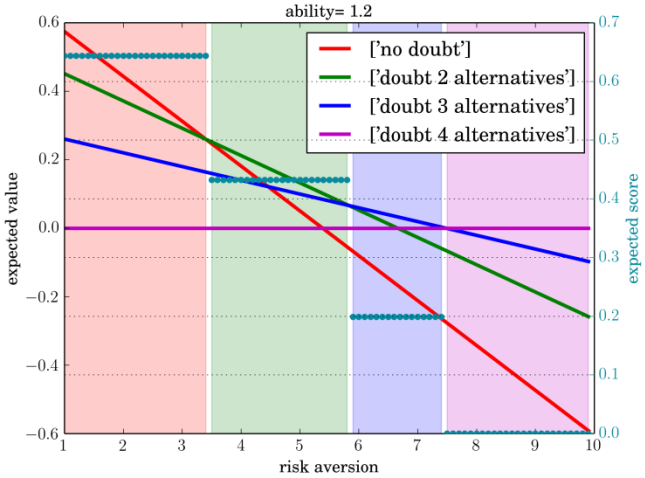

Fig. 3: Example of step 2 for EMA: the value of the different answer patterns (left y-axis) for ability=1.2 (4 alternatives). According to prospect theory, the student will choose the answer pattern with maximum value. This maximum value answer is shown by the background colour. The cyan dots (right y-axis), the corresponding score. The more risk-averse a student, the more his answering pattern will include bout, and the lower the expected score of that student.

The combination of step 1 and step 2 allows obtaining the maximum value answer, the expected score, and the expected variance on the score for a particular question in function of the risk aversion and ability. Fig. 4 shows an example of the results for EMA.

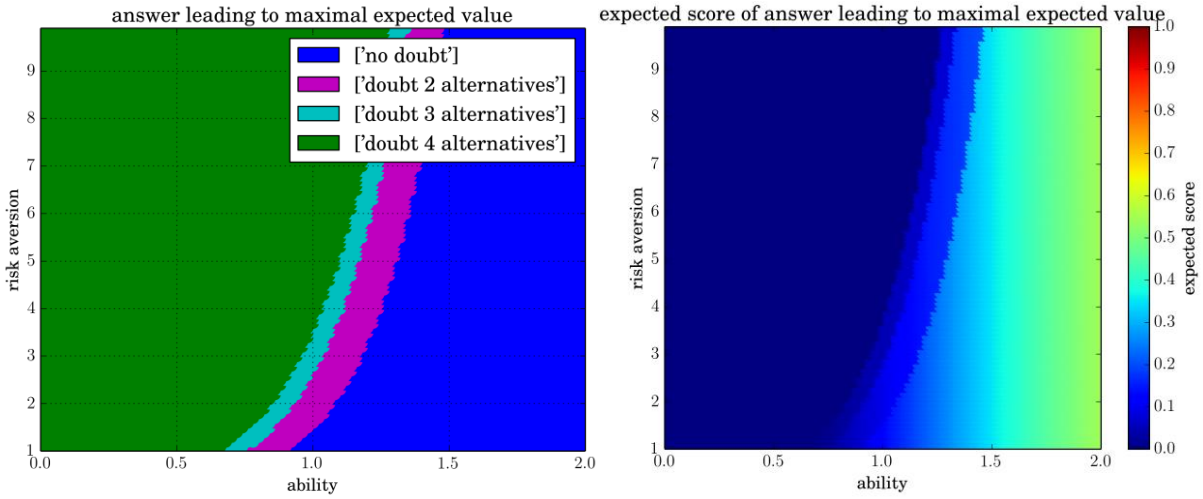

Fig. 4: Result of theoretical analysis of one MPQ for EMA showing answer pattern (top left) and expected score (right) in function of ability and risk-aversion (4 alternatives).

Findings

The theoretical analysis confirms that SS is independent on risk-aversion. However, the expected variance is higher than NM and EMA, especially for low and medium ability students (*Fig*. 5). This decreases the reliability of the SS scoring method.
Furthermore, the analysis shows that students will actually use the opportunity to indicate their doubt when answering questions (*Fig. 4*, left).

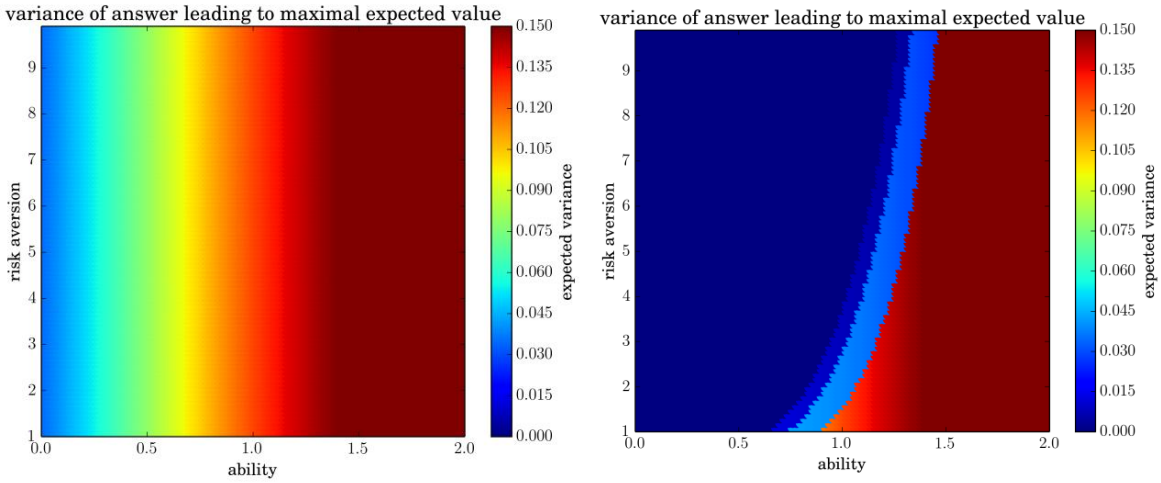

Fig. 5: Variance of maximum value answer for SS (left) and EMA (right).

Finally, the expected score and variance of NM and EMA are very similar; they only differ in the small region where students will indicate doubt (*Fig. 6*). In the regions where the students are expected to indicate doubt, EMA reduces the effect of risk-aversion on the score and reduces the variance. Therefore, EMA is a valuable

alternative for NM that has a small impact on the overall expected score, but decreases the dependency of risk-aversion and increases the reliability of the exam.

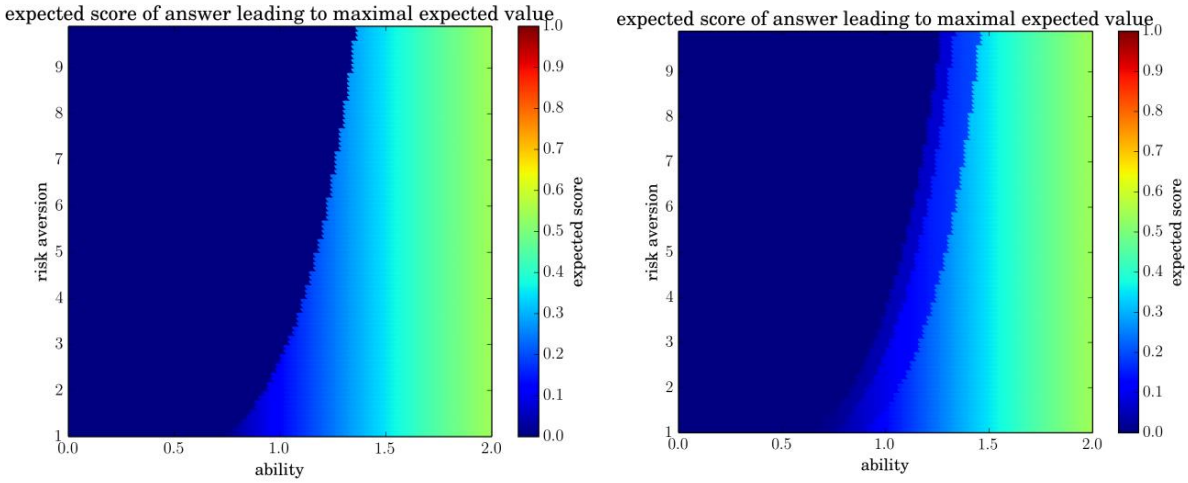

Fig. 6: Expected score of maximum value answer for NM (left) and EMA (right).

3.2 Empirical analysis

The **empirical analysis** gathers evidence from two exams of first-year engineering students with elimination marking. It investigates how students use elimination marking to indicate doubt, and the influence of gender. Finally a questionnaire probing for students' opinions on EMA and NM was performed. All students had previous experiences with NM.

Findings

Table 4 presents the results of empirical analysis are for the two exams. Almost all students express partial knowledge on at least one of the questions (doubt in table). Between 1/10th and 1/4th of the answers of the questions shows an answer pattern with doubt (without blanks). The gender difference is statistically significant ($p<.001$), even when conditioned on the ability (measured as the average percentage of all course grades, $p<.01$), but is strangely reversed between the two exams. The overall exam scores do not show a statistically significant difference regarding gender. Moreover, the comparison with previous academic years shows that, as was predicted [5], the overall exam statistics (average, median, percentage passed) did not change. With respect to previous years, the amount of partial misconception and full knowledge decreased, while the amount of measured partial knowledge increased. This indicates that students use the options to express partial knowledge instead of guessing.

In the questionnaire students (N=134) confirm that the instructions of EMA are clear (87%, 13%, 0%, 0%, 0%), where the reported percentages are according to a 5-point Likert scale (strongly agree, agree, undecided, disagree, strongly disagree). They find NM more difficult than EMA (8%, 47%, 28%, 13%, 4%). Furthermore, they report higher stress levels with NM than with EMA (12%, 49%, 18%, 17%, 4%). EMA is however considered more time consuming than NM (12%, 49%, 18%, 17%, 4%). Most importantly, they overall prefer EMA over NM (33%, 45%, 13%, 5%, 4%).

Table 4: Results empirical analyses Electrical Networks and Philosophy: different knowledge levels (see *Table 2*) expressed in % of students and % of questions, both for male and female students

				FK	PK1	PK2	NK	PM1	PM2	TM	doubt = (PK1+PK2+ PM2+TM)
electrical networks (437 students 26 questions)	**% students**	**total**	100,0	99,8	66,6	28,8	88,1	95,9	36,4	6,9	79,4
		male	82,8	99,7	64,4	26,5	87,0	97,0	34,3	5,5	77,3
		female	15,3	100	77,6	38,8	92,5	89,6	43,3	11,9	89,6
	% questions	**total**	100,0	52,5	5,8	1,8	19,7	17,1	2,7	0,3	10,7
		male	82,8	52,7	5,6	1,6	20,0	17,2	2,4	0,3	10,0
		female	15,3	54,1	6,4	2,8	17,5	15,0	3,7	0,5	13,4
philosophy (454 students 30 questions)	**% students**	**total**	100,0	99,8	95,6	46,3	56,6	97,6	55,1	7,9	96,7
		male	84,1	99,7	97,1	48,4	56,5	97,6	55,8	8,4	97,9
		female	15,9	100	87,5	34,7	56,9	97,2	51,4	5,6	90,3
	% questions	**total**	100	53,9	15,3	2,8	4,9	19,2	3,6	0,3	22,0
		male	84,1	53,2	16,0	2,9	4,9	19,1	3,7	0,3	22,8
		female	15,9	57,6	12,0	2,4	5,3	19,5	3,0	0,2	17,6

4 CONCLUSIONS AND RECOMMENDATIONS FOR FURTHER WORK

The paper has two contributions. Firstly, it introduces and uses a theoretical framework that combines statistics, economic, and educational models to study the effect of risk-aversion and ability for different scoring methods: NM, SS, and EMA. Secondly, it shows using an empirical analysis presenting results of EMA on examination for first-year engineering bachelor students. The theoretical and empirical analyses show that EMA is a valid alternative for NM, and is preferred by students as it improves satisfaction and reduces anxiety.

Further work will concentrate on further developing the theoretical analysis and in particular the combination of economic models and educational models. So far, the analysis is done for single questions. Therefore, extending the analysis to entire MPQ exams is a priority, while this is an obvious extension. Additionally, alternative scoring methods for MPQ exams can be included in the analysis. The empirical analysis could be further strengthened by including more direct comparisons of different scoring methods on the same MPQ exams with similar student populations. This is however difficult to obtain from an ethical point-of-view.

5 ACKNOWLEDGMENTS

We gratefully acknowledge the support of the KU Leuven project OWP IMP2015/08 and LESEC- the Leuven Engineering and Science Education Centre.

REFERENCES

[1] E. Lesage, M. Valcke, and E. Sabbe, “Scoring methods for multiple choice assessment in higher education - Is it still a matter of number right scoring or negative marking?,” *Studies in Educational Evaluation*, vol. 39, no. 3. pp. 188–193, 2013.

[2] A. R. Hakstian and W. Kansup, “A Comparison of Several Methods of Assessing Partial Knowledge in Multiple-Choice Tests: II. Testing Procedures,” *J. Educ. Behav. Stat.*, vol. 12, no. 4, pp. 231–239, 1975.

[3] a. Ben-Simon, D. V. Budescu, and B. Nevo, “A Comparative Study of

Measures of Partial Knowledge in Multiple-Choice Tests," *Appl. Psychol. Meas.*, vol. 21, no. 1, pp. 65–88, 1997.

[4] W. Kansup and A. R. Hakstian, "A comparison of several methods of assessing partial knowledge in multiple-choice tests: I. Scoring procedures," *J. Educ. Meas.*, vol. 12, no. 4, pp. 219–230, Dec. 1975.

[5] T. De Laet, J. Vanderoost, R. Callens, and J. Vandewalle, "How to remove the gender bias in multiple choice assessments in engineering education ?," in *Proceedings of the 43th Annual SEFI conference*, 2015, pp. 2–9.

[6] Y. Bereby-Meyer, J. Meyer, and O. M. Flascher, "Prospect Theory Analysis of Guessing in Multiple Choice Tests," *J. Behav. Decis. Mak.*, vol. 15, no. 4, pp. 313–327, 2002.

[7] G. Ben-Shakhar and Y. Sinai, "Gender Differences in Multiple-Choice Tests: The Role of Differential Guessing Tendencies," *Source J. Educ. Meas. J. Educ. Meas. Spring*, vol. 28, no. 1, pp. 23–35, 1991.

[8] K. Baldiga, "Gender Differences in Willingness To Guess," *Manage. Sci.*, vol. 60, no. 2, pp. 434–448, Feb. 2014.

[9] A. E. Bond, O. Bodger, D. O. F. Skibinski, D. H. Jones, C. J. Restall, E. Dudley, and G. van Keulen, "Negatively-Marked MCQ Assessments That Reward Partial Knowledge Do Not Introduce Gender Bias Yet Increase Student Performance and Satisfaction and Reduce Anxiety," *PLoS One*, vol. 8, no. 2, 2013.

[10] Y. Bereby-Meyer, J. Meyer, and D. V. Budescu, "Decision making under internal uncertainty: The case of multiple-choice tests with different scoring rules," *Acta Psychol. (Amst).*, vol. 112, no. 2, pp. 207–220, 2003.

[11] C. H. Coombs, J. E. Milholland, and F. B. Womer, "The assessment of partial knowledge," *Educ. Psychol. Meas.*, vol. 16, no. 1, pp. 13–37, 1956.

[12] J. C. Arnold and P. L. Arnold, "On Scoring Multiple Choice Exams Allowing for Partial Knowledge," *Source J. Exp. Educ. J. Exp. Educ.*, vol. 39, no. 1, pp. 8–13, 1970.

[13] H. Fennema and P. Wakker, "Original and cumulative prospect theory: a discussion of empirical differences," *J. Behav. Decis. Mak.*, vol. 10, no. 1, pp. 53–64, 1997.

[14] N. C. Barberis, "Thirty Years of Prospect Theory in Economics," 2012.

[15] R. B. Frary, "Partial-Credit Scoring Methods for Multiple-Choice Tests," *Applied Measurement in Education*, vol. 2, no. 1. pp. 79–96, 1989.

[16] G. N. Masters, "A rasch model for partial credit scoring," *Psychometrika*, vol. 47, no. 2, pp. 149–174, 1982.