CIID
I CONGRESO INTERNACIONAL DE INNOVACIÓN DOCENTE
20 y 21 de FEBRERO DE 2014
MURCIA - CARTAGENA (ESPAÑA)

"THE INFLUENCE OF RISK PERCEPTION AND SCRAMBLING ON STUDENTS PERFORMANCE"

Autor/res/ras: Antonio J. Carrasco Hernández, M. Encarnación Lucas Pérez, Alicia Rubio Bañón y Gregorio Sánchez Marín

Institución u Organismo al que pertenecen: Universidad de Murcia

Indique uno o varios de los seis temas de Interés: (Marque con una (x))

( ) Enseñanza bilingüe e internacionalización

( ) Movilidad, equipos colaborativos y sistemas de coordinación

( ) Experiencias de innovación apoyadas en el uso de TIC. Nuevos escenarios tecnológicos para la enseñanza y el aprendizaje.

( ) Nuevos modelos de enseñanza y metodologías innovadoras. Experiencias de aprendizaje flexible. Acción tutorial.

( ) Organización escolar. Atención a la diversidad.

(X) Políticas educativas y reformas en enseñanza superior. Sistemas de evaluación. Calidad y docencia.

Idioma en el que se va a realizar la defensa: (Marque con una (x))

( ) Español **(X)** Inglés

Resumen.

En este trabajo se analiza si existe una variación en las calificaciones de los estudiantes como consecuencia del orden de las preguntas y de las penalizaciones de las respuestas incorrectas. Se han creado dos escenarios que incluyen diferentes exámenes según diferente ordenación de contenidos y riesgo percibido. Utilizando una muestra de 764 exámenes de primer grado de Relaciones Laborales y Gestión de Recursos Humanos de la Universidad de Murcia, los resultados indican que hay diferencias en el procedimiento de respuesta que afecta a la nota final, cuando las percepciones de riesgo de los estudiantes son diferentes. Por otra parte, el efecto de orden también influye en el procedimiento de respuesta, influyendo asimismo en las calificaciones finales de los alumnos en función del escenario de riesgo al que los estudiantes se enfrentan.

Palabras Claves: Examen test, ordenación de contenidos, penalización de respuestas, riesgo percibido

Abstract

This paper analyzes if there is a variation in students’ marks due to the type of multiple choice exam in relation to the order of the questions and the penalty of the incorrect answers. We create two scenarios including exams sorted both by different order content and by different level of risk borne. Using a sample of 764 multiple choice exams of Industrial Relations and Human Resource Management Grades of University of Murcia (Spain), results lead us to affirm that there are differences in the answer’s procedure, affecting the final marks when risk perceptions of the students are different. Moreover, the order effect also influences the response procedure and its effects on the final marks of the students depending on the risk scenario the students are facing.

Keywords: Multiple choice exams, order content, penalty responses, risk perception.

1. Introduction

The development of new information and communication technologies (ICTs) has facilitated many of the roles of teachers making more efficient their daily tasks. Access to ICTs has not only improved the means of teaching in the classroom but also has promoted the existence of other communication and assessment channels. In that way, ICTs have allowed performing multiple choice exams with immediate feedback for the students. Since one of the most important processes of interaction between student and teacher is the assessment of knowledge, skills and attitudes acquired in classroom, analyze the results of multiple choice exams has been of great importance in the literature of education (Doerner and Calhoun, 2009; Sue, 2009). The debate that has emerged around this type of testing is that the level of difficulty and the results obtained by the student is highly determined not only by the level of complexity of the questions, but also by the design of parameters –number of questions, ordination of questions, penalty for incorrect answers,...- affecting the student response process (Bresnock et al., 1989; Sue, 2006).

Some investigations have shown that the variation in students’ marks is strongly affected by the type of exam –in relation to the order of the questions- (Taub and Bell, 1975; Carlson and Ostrosky, 1992; Doerner and Calhoun, 2009) and the penalty of the incorrect answers –in reference to the level of risk borne by the students- (Carrasco et al., 2013). Taking into account these evidences, in this study we focus on analyzing these two parameters of multiple choice exams, creating two scenarios that include exams sorted both by different order content and by different level of risk borne. Regarding risk, although both scenarios penalize incorrect answers in the same proportion, there are distinct students’ perceptions: in the first scenario students know that their incorrect answers are penalized while, in the second one, students do not know that their incorrect answers are penalized. Considering those different scenarios, in this paper we would like to answer the following questions: Does penalization and content order affect the responses given and the qualifications obtained by students? Which are the most affecting variables in terms of students' marks?

The study was conducted on a sample of 764 first grades' multiple choice exams of Industrial Relations and Human Resource Management. Our general findings show that there are differences in the answer's procedure, affecting the final marks when risk perceptions of the students are different. Students who thought they would not be penalized in the multiple choice exam were obtained better marks and a higher rate of correct answers, as well as a lower rate of incorrect answers. Moreover, the order effect also influences the response procedure and its effects on the final marks of the students depending on the risk scenario they are facing.

To reach our objectives first, we review the main studies investigating these issues. Second, we describe the sample and the measurement of variables, explaining analysis and results obtained. Finally, we discuss the conclusions reached.

2. Theory and hypotheses

There is abundant literature corroborating that the use of multiple choice exams increases potential for cheating. One method of counteracting this fact is to use multiple forms of the same exam by scrambling test questions. In that way Chatterjee (2013) states that many early papers theoretically studied the relation between distributions of items' difficulties and the reliability of exams. Other studies have examined the estimation methods and confidence intervals, the role of chance on test validity, the correlations between tests and their relation to test reliability, the indexes of cheating, the role of latent traits including gender-based differences in test taking, etc. In all cases, the evaluation of multiple choice exams has been a combination of several content order's designs with different score responses penalization's scenarios, with the aim of obtaining an adequate design of multiple-choice questions.

The scramble questions method usually produces one content ordered exam that matches the order in which the material was presented in class and another one that has a random scrambled order. Although research has study this fact there has been no consensus about the real effect of scrambling questions, previous studies do affirm that students may perform better on a content-ordered exam (Sue, 2009). The students who took the form which questions were randomly arranged have more test anxiety than their classmates and they also lose concentration (Taub and Bell, 1975). However other studies showed opposite results (Bresnock et al., 1989; Gohmann and Spector, 1989), highlighting that to analyze the real effect of scrambling new mediating variables are needed as, for example, the general qualifications of students, their experiences in education career, their risk aversion levels, and so on.

In that way, recent studies identified different behavior of some groups when solving an ordered or a randomized test exam, suggesting that test scrambling really affects students behavior depending risk aversion (Marín and Rosa-Garcia, 2012).

There are mainly two possibilities to introduce risk in a multiple choice exam. The first one is based on a conventional number right scoring method (Bereby-Meyer et al., 2002). Correct answers are scored with a positive value, incorrect answers and absent or omitted answers with a value of zero, using a dummy to score the test (Kurz, 1999). This type of correction encourages students to pass the exam answering random questions, because there is no penalty. Those skilled students could pass the exam trying their luck and increasing the number of questions answered (Abu-Sayf, 1979; Choppin, 1988; Budescu and Bar-Hillel, 1991; Kubinguer et al., 2010). All of which, increased discussion of how to distinguish between those students who have actually acquired the right skills and they are able to pass the course and those who try without luck to be able to pass the course. To help avoid this problem, many methods have been suggested correction to try to minimize as far as possible the chance to pass an exam, for example, calculating the score based on certain algorithmic models that can avoid some of these problems.

The second scoring method introduces risk. Among the most used today are called “rights minus wrongs” correcting model (Kurz, 1999). In this case students are penalized for their incorrect responses. As that students acknowledge they will lose marks for incorrect answers they are discouraged to guess, and this is expected to increase test reliability and validity because a score is obtained with less likelihood of bias that can more reliably assess the acquisition of knowledge by students.

The discussion of which is the most reliable way of scoring is still open. When students are not penalized there is a greater likelihood that the score does not really reflect whether students have acquired the knowledge they should because they can answer the questions using the random (Choppin, 1988; Kurz, 1999). However, if the correct answers can penalize, the decision can answer questions more or less dependent on the type of risk aversion that students have. Thus, a more risk adverse students tend to answer fewer questions, while those who are less risk averse and not valued in the penalty decision, they will answer more questions in terms also of chance and probability of success (Albanese, 1988; Angoff, 1989), valuing the decision may, in this case, the strategy of most decision and risk assumed, really the knowledge acquired by the student (Choppin, 1988; Kurz, 1999).

According to these arguments, and considering that the student answers can be seen influenced by the ordination of answers and by their degree of risk aversion when wrong answers are penalized, we propose the following research hypothesis:

> ***H1**: Students who perform an ordered questions examination (type 01)do better than those performing a not ordered questions examination (type 02).*
>
> ***H2**: Students who perform a low risk perceiving examination do better than those performing a high risk perceiving examination.*
>
> ***H3:** The response procedure (right, wrong and blank questions) of students doing type 01 differs from those doing the type 02, as well as of students who perceive low risk from those who perceive high risk.*

3. Methodology

3.1. Sample and data

In order to empirically test the effect of scrambling the content order of multiple-choice questions on a student’s performance and exploring the differences in outputs, we use a sample of 764 type 01 and 02’s exams done by undergraduate students from the University of Murcia (Spain) belonging to first grade of Industrial Relations and Human Resource Management of the 2012-13 academic year.

3.2. Variables and scales

We used dependent, independent and moderating variables. The dependent variable is student performance, measures in three ways: (1) as a quantitative variable in a 0-10 points' scale, which indicate the score obtained by the student in the exam; (2) as a dummy variable indicating if the students obtain a score below or above average; (3) as a qualitative variable indicating the procedure of answer, measures as the number of correct, incorrect and blank responses doing by the student in the exam. Our two independent variables are: the content ordination, measures through a dummy variable indicating type 01 exam (with ordered questions) and type 02 (with not ordered questions), and risk perception, measures as a dummy variable indicating high risk (when student knows to be penalized for wrong questions) and low risk (when the student thinks he is not penalized, but if he truly penalized). We also have used two segmentations, which lead us to generate two moderating variables: (1) experience of the students, proxies by two variables: the age of students (more or less of 19 years old) and the number of exam' call student facing (first, second or more); (2) risk taking by the student doing the exam, proxies by the genre of student as a qualitative variable indicating male or female.

4. Results

Our general results indicate that there are significant differences between the scores obtained by students in the exams with a penalty for incorrect answers and exams without penalty. In order to compare the results and discuss the efficiency of the students in the exams, the test scores without penalty to the wrong questions were calculated according to the same penalties as those applied in the tests with a penalty for incorrect answers.

To better understand the behavior of students in case of penalty for incorrect answers, the response process and the scores of ordered content exams (type 01) and disordered content exams (type 02) are examined. Thus, as table 1 show, there are not significant differences between the type 01 and 02 exams performance in the exams with penalty for incorrect answers, but there are significant differences between the type 01 and 02 exams performance in the exams without a penalty for incorrect answers. The no penalization of multiple choice exams affects the results obtained by the students and the response procedure.

Specifically, neither differences in scores on both multiple choice exams –type 01 and type 02- have been found nor differences in students with higher marks (those above the average score) compared to students with lower marks (those below the average score) in exams with penalty for incorrect answers. We only found differences in case of students with lower marks in the case of type 02 exams, in which students left more blank answers and have fewer wrong answers. Hence, our results lead us to affirm that although no differences in scores on both type 01 and 02 multiple choice exams have been found, there are differences in the procedure of response, in spite of this procedure does not affect the final performance.

In exams without penalty for incorrect answers, we found differences in case of students with lower grades: specifically, type 01 exams' students obtained worse scores, more wrong answers and have fewer right answers. Hence, our results indicate that although type 01 questions follow the same order than the class syllabus, students do worse. One possible explanation is that decreasing the degree of concentration of the students to feel more confident, having questions sorted by class syllabus and not be penalized for incorrect answers.

Trying to explain and understand the origin of such differences, we have segmented the study sample in terms of: (1) student experience in relation to test examination, and (2) the student risk taking doing the exam.

In the experience dimension, first we compare older students (those above the average age) to younger students (those below the average age). As can be seen in table 1, in exams with penalty for incorrect answers, results do not show differences in type 01 and 02 performance of younger students, but show differences in case of older students: students doing type 02 exams left more blank answers. Second, we compare the student's experience facing the exam (first call vs. second or more call), obtaining no differences in type 01 and 02 performance of second call or more students. However, we found significant differences in performance exams between first call students: type 02 exams' students left more blank answers. In the risk taking dimension, results show no differences in type 01 and 02 performance exams in female group of students. In case of male, there are differences: in 02 type exams' students left more blank answers and have less wrong answers, not affecting final performance.

In exams without penalty for incorrect answers, the results show differences in type 01 and 02 performance exams of older students: type 02 exams' student obtain better scores, more right answers and less wrong answers. In case of younger students, the type 02 obtain better scores and more right answers. Second, we have compared the student's experience facing the exam (first call vs. second or more call), obtaining no differences in type 01 and 02 performance exams of first call students. However, we found significant differences in performance exams between second call or more students: type 02 exams' students obtain more right answers, less wrong answers and better scores. In the risk taking dimension by genre, results show no differences in type 01 and 02 performance exams in male group of students. In case of female, there are differences: in 02 type exams' students obtain better scores, more right answers and less wrong answers. Again we find that the gender segmentation offers significant differences in the response processes of students

Table 1. Results of analysis of variance (ANOVA)

		% right answer	% wrong answer	% blank answer	Scores
General comparison by risk (penalty)					
	Risk (penalty)	44,66%***	35,10%***	20,23%***	3,34***
	No risk	56,70%***	42,65%***	0,65%***	4,26***

		Risk (penalty)				No risk (no penalty)			
	Exam Type	% right answer	% wrong answer	% blank answer	Scores	% right answer	% wrong answer	% blank answer	Scores
General comparison by risk (penalty) and exam type									
	01	45,21%	36,40%	18,38%**	3,34	52,54%***	46,72%***	0,74%	3,72***
	02	44,08%	33,79%	22,12%**	3,34	58,52%***	40,86%***	0,61%	4,49***
Best grade students	01	56,97%	22,67%	20,35%	4,94	69,55%	29,90%	0,55%	5,95
	02	56,38%	22,67%	20,49%	4,92	69,05%	30,56%	0,38	5,88
Worse grade students	01	32,94%	50,74%**	16,32%***	1,66	38,41%**	60,67%**	0,91%	1,87**
	02	30,38%	45,76%**	23,86%***	1,62	41,86%**	57,16%**	0,98%	2,30**
Students experience									
A. Students average age									
Older students	01	45,89%	37,85%	16,25%*	3,34	53,56%**	45,62%**	0,81%	3,38**
	02	41,25%	34,68%	24,06%*	2,98	59,46%**	39,97%**	0,56%	4,61**
Younger students	01	45,04%	36,03%	18,91%	3,33	49,92%*	49,48%	0,59%	3,37*
	02	44,95%	33,52%	21,52	3,44	55,78%*	43,43%	0,77%	4,16*
B. Students number of exam' call facing									
First call	01	42,14%	32,98%	24,88%**	3,13	45,48%	54,26%	0,26%	2,79
	02	40,35%	32,09%	27,56%**	2,99	51,62%	47,95%	0,43%	3,58
Second call or more	01	46,55%	37,89%	15,57%	3,42	59,04%**	39,76%**	1,20%	4,59**
	02	45,80%	34,57%	19,63%	3,49	62,72%**	36,55%**	0,72%	5,05**
Students risk taking by genre									
Women	01	46,57%	34,10%	19,34%	3,56	51,71%***	47,29%***	1,00%	3,64***
	02	42,85%	35,32%	21,84%	3,20	58,76%***	40,51%***	0,73%	4,54***
Men	01	43,21%	39,82%***	16,96%**	3,00	53,66%	45,95%	0,39%	3,85
	02	45,78%	31,72%***	22,50%**	3,52	58,21%	41,33%	0,46%	4,44

* p<0,1; **p<0,05; ***p<0,01

Finally, test results improve when the wrong exam answers are not penalized, more when questions are randomly disordered to achieve greater concentration of the student. It would therefore be advisable to establish test exams without penalty for wrong answers and increase the cut-off to pass the test if student outcomes are formed in conjunction with other scores.

Thus, we reject hypothesis 1 since students who perform an examination of the type 02 do better than those doing the type 01 whereas we accept hypothesis 2 since students who perceive low risk do better than those who perceive high risk. Furthermore, we accept hypothesis 3 because the response procedure (right questions, erroneous and white) of students taking the test type 01 differs from those doing the exam type 02 as well as differs from those who perceive low risk comparing to those who perceive high risk.

5. Conclusions

The technological revolution has not only meant a change in traditional teaching methods, but has helped teachers in various aspects. One of the main foundations of teaching is the evaluation of the subjects. New ICTs have allowed teachers to develop different multiple choice exams types, alternating ordered and random questions as well as high and low level of penalization. In fact, several studies have tried to analyze whether the type of multiple choice exam, considering the grade of randomization and penalization can positively or negatively affect the performance and the response procedure of the students. However, mixed results have been reached by literature (Abu-Sayf, 1979; Albanese, 1988; Angoff, 1989; Bereby-Meyer et al., 2002; Bresnock et al., 1989; Budescu and Bar-Hillel, 1991; Chatterjee, 2013; Choppin, 1988; Gohmann and Spector, 1989; Kubinguer et al., 2010; Kurz, 1999; Marín and Rosa-Garcia, 2012; Taub and Bell, 1975).

With the aim of bring some lights to this controversial matter, our study has tried to know if penalization and content order affect the responses given and the qualifications obtained by students and, in affirmative case, which are the most affecting variables in terms of students' marks. For this purpose, we conducted some statistical analyses based on a sample of 764 exams of Industrial Relations and Human Resource Management Grade.

Our general findings show that there are differences in the answer’s procedure, affecting the final marks when risk perceptions of the students are different. Students who thought they would not be penalized in the multiple choice exam were obtained better grades and a higher rate of right questions, as well as a lower rate of incorrect answers. Moreover, the order effect also influences the response procedure and its effects on the final marks of the students depending on the risk scenario they are facing. These evidences agree with some other recent investigations (Marin and Rosa-García, 2012). Also, we have found that a more risk adverse students tend to answer fewer questions, while those who are less risk averse and not valued in the penalty decision; they will answer more questions in terms also of chance and probability of success, as other earlier studies said (Albanese, 1988; Angoff, 1989). With exams without penalty for wrong answers, the student scores are better, and that supposes an advantage for the student.

The question is whether better results are due to chance or to the degree of knowledge acquired by the student as said Choppin (1988) and Kurz (1999). That is why there is need for further research. Specifically, future studies should take into account If there are different procedures in response rates is useful for further examination and different in order to see what might be the causes that lead students to fail in another type of test, helping to decrease the biases that can adversely affect the performance of the multiple choice exams.

6. References

Abu-Sayf, F.K. (1979). The scoring of multiple choice tests: A closer look. *Educational Technology*, 19, 5–15.

Albanese, M.A. (1988). The projected impact of the correction for guessing on individual scores. *Journal of Educational Measurement*, 25, 149-157.

Angoff, W.H. (1989). Does guessing really help? Journal of Educational Measurement, 26, 323-336.

Bereby-Meyer, Y.; Meyer, Y.; Flascher, O. M. (2002). Prospect theory analysis of guessing in multiple choice tests. *Journal of Behavioral Decision Making,* 15, 313–327.

Bresnock, A.E.; Graves, P.E.; White, N. (1989). Multiple-Choice Testing: Question and Response Position. *Journal of Economic Education*, 20, 239-245.

Budescu, D.; Bar-Hillel, M. (1993). To guess or not to guess: a decision-theoretic view of formula scoring. *Journal of Educational Measurement,* 30, 277–291.

Carlson, J.L.; Ostrosky, A.L. (1992). Item Sequence and Student Performance on Multiple-Choice Exams: Further Evidence. *Journal of Economic Education*, Summer, 232-235.

Carrasco, A; Lucas, M.E.; Marín, C.; Rubio, A.; Sánchez, G. (2013). Does content order affect performance on multiple choice exams? International Technology, Education and Development (INTED) Conference. Valencia, Spain.

Choppin, B.H. (1988). Correction for guessing. In J.P. Keeves (Ed.), Educational research, methodology, and measurement: An international handbook, 384–386. Oxford: Pergamon Press.

Doerner, W.M.; Calhoun, J.P. (2009). The Impact of the Order of Test Questions in Introductory Economics. Economics Educator: Courses, Cases & Teaching, April 20, Economics Research Network (ERN) Working Paper.

Gohmann, S.F.; Spector, L.C. (1989). Test Scrambling and Student Performance. *Journal of Economic Education*, 20, 235-238.

Kubinger, K.D.; Holocher-Ertl, S.; Reif, M.; Hohensinn, C.; Frebort, M. (2010). On minimizing guessing effects on multiple-choice items: Superiority of a two solutions and three distractors item format to a one solution and five distractors item format. *International Journal of Selection and Assessment,* 18, 111-115.

Kurz, T.B. (1999). A review of scoring algorithms for multiple-choice tests. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX.

Marín, C. and Rosa-García, A. (2012). "Gender Bias in Risk Aversion: Evidence from Multiple Choice Exams". MPRA Paper nº 39987, Julio 2012. Online at http://mpra.ub.inimuenchen.de/39987/."

Sue, D.L. (2006). The Effect of Test Scrambling on Student Performance. NBEA Annual Conference Proceedings.

Sue, D.L. (2009). The effect of scrambling test questions on student performance in a small class setting. *Journal for Economic Educators*, 9, 32-41.

Taub, A.J.; Bell, E.B. (1975). A Bias in Scores on Multiple-Form Exams. *Journal of Economic Education*, 7, 58-59.