Educational Measurement
ISSUES AND PRACTICE
NCME national council on measurement in education

Educational Measurement: Issues and Practice
Fall 2019, Vol. 38, No. 3, pp. 20–32

The Value of Choice: An Experiment Using Multiple-Choice Tests

Henry Aray, *Universidad de Granada,* and Luis Pedauga, *University of Leon*

This article presents a novel experimental methodology in which groups of students were offered the option to choose between two equivalent scoring rules to assess a multiple-choice test. The effect of choosing the scoring rule on marks is tested. Two major contributions arise from this research. First, it contributes to the literature on the value of choice. Second, it also contributes to the literature on the educational measurement of knowledge. The results suggest that choice could positively affect students′ scores. However, students need to learn to choose the assessment method. Moreover, women seem to obtain greater benefits from the option of choosing the scoring rule.

Keywords: choice, field experiment, multiple-choice questions

It is widely known that freedom of choice has value. This issue has received much attention across different disciplines, especially economics and psychology,[1] which recognize the value of choice both intrinsically and instrumentally. In fact, Iyengar and Lepper (1999) pointed out that American psychologists have contended that providing choice will increase the individual's sense of personal control and feelings of intrinsic motivation, which have been associated with numerous physical and psychological benefits. Dowding and John (2009) define choice as being instrumentally valuable in the sense that increasing choice in public services brings welfare gains through efficiency by the signals that choice gives to providers. They converge to the same conclusion as psychologists that choice enhances individual autonomy. They also pointed out that a sense of any intrinsic value can be further justified instrumentally.

According to Usher, Elhalal, and McClelland (2008), a theory of choice is paramount in all domains of cognition requiring behavioral output. Experimental psychology and neuroscience disciplines focus on perceptual choice, while economics and social sciences focus on preferential choice. The new field of neuroeconomics has bridged these disciplines, as it aims at understanding the principles that underlie value-based decisions and the neural mechanisms through which these principles are expressed in behavior. Thus, neurophysiological studies of value-based decisions and neurocomputational models of preferential choice have been developed by Glimcher (2004) and Sugrue, Corrado, and Newsome (2004, 2005).

In this article, we aim to find the value of choice using multiple-choice tests. We propose a novel approach based on an experimental methodology that empowers students with the option to choose between two equivalent formula scoring rules to asses multiple-choice questions (MCQ). We wonder if the ability to choose the scoring rule leads to higher grades. The contribution of this article is twofold. First, it contributes to the literature on the value of choice since we quantify how much the option to choose the scoring rule adds to the students' scores. Second, it also contributes to the literature on the educational measurement of knowledge because our proposal could be useful in increasing the correlation between knowledge and scores, which is the final objective of the examiner. Students are expected to improve their scores through the power of choice. Thus, in the experiment that we conducted, subjects are not considered passive decision makers who follow a behavioral rule, as usually assumed, but rather as individuals who choose the formula scoring rule based on their individual characteristics.

Our proposal can be framed in the analysis of the rational behavior of students, which has been studied by Espinosa and Gardeazabal (2010, 2013), who claim that the analysis of subjects' behavior in MCQ provides a scenario for studying the relationship between risk attitudes and knowledge.

MCQ examinations, regardless of the degree of difficulty, are among the most objective tools for assessing the acquisition of skills on a subject, and they are among the most widespread methods of knowledge assessment worldwide. The main advantage of MCQ examinations is that a larger number of questions can be asked, which allows to for greater coverage of the content area, regardless of the students' writing speed. Bacon (2003) adds that even though high levels of knowledge are measured, MCQ is an appropriate measurement tool. Specifically, in the field of economics, Siegfried, Saunders, Stinar, and Zhang (1996) and Bredon (2003) have highlighted their usefulness. They claim that the main advantage of MCQ over other types of exams, such as oral or constructed responses, is that it precludes measurement errors introduced by the grader. Its main disadvantage, according to the above authors, is that it might encourage guessing and random responses, which could reduce their reliability in measuring students' knowledge on a particular subject. In

Henry Aray, Department of Economics, Facultad de Ciencias Económicas y Empresariales, Universidad de Granada, Campus de la Cartuja S/N, 18011, Granada, Spain; haray@ugr.es. Luis Pedauga, University of Leon, Leon, Spain; lpeds@unileon.es.

order to overcome this, examiners often use the well-known correction for guessing formula, a rule that penalizes wrong answers in order to discourage questions answered randomly. However, Espinosa and Gardeazabal (2010, 2013) consider that the introduction of such a penalty leads the student to a problem of decision making under uncertainty. That is, in the case of not knowing the answer to a question, the option of responding can be seen as a lottery in which the student chooses between not responding (not playing) and obtaining a secure payment (usually zero) or responding (playing) and obtaining a nonzero payment, positive with some probability p, or negative with the complementary probability $(1 - p)$. The option to answer the question and thus to play the lottery depends, among other factors, on the student's risk aversion, while the probability of obtaining a correct answer depends mainly on the student's knowledge of the subject matter.

Moreover, Espinosa and Gardeazabal (2010, 2013) noted that the correction for guessing formula could discourage random answers in a nonuniform way among students, that is, students with high risk aversion are more easily discouraged, which introduces bias toward risk-neutral or risk-loving students. Consequently, with equal total or partial knowledge, risk-averse students could obtain lower scores on average than risk-neutral or risk-loving students. In addition, they suggest that risk attitudes are different across domains. Along the same line, Prieto and Delgado (1999) pointed out that if risk aversion is correlated with gender, knowledge, social group, or other characteristics of the students, the penalty adversely discriminates against that group of students. Precisely because of this, the Educational Testing Service (ETS) recently stopped using the correction for guessing formula in the Scholastic Assessment Test (SAT). One of the persistent statistical results from the SAT was the correlation between high income and high test scores. The ETS has redesigned the SAT to reinforce the skills and evidence-based thinking that students should be learning in the classroom instead of test-taking tricks and strategies. Therefore, it is intended to be fairer and more equitable. Moreover, Bolger and Kellaghan (1990), Ben-Shakhar and Sinai (1991), and Beller and Gafni (2000) have found that, on average, males obtain higher scores than females in MCQ exams.

Typically, the literature alternatively suggests using number-right scoring (no penalty), which induces students to answer all questions, encouraging even guessing and offsetting the bias due to risk aversion heterogeneity. However, Espinosa and Gardeazabal (2010, 2013) suggest that the elimination of the penalty using the number-right scoring rule, although it allows risk-neutral and risk-averse students to get the same expected score which can be interpreted as a lack of discrimination against risk-averse students, does not eliminate the variability of actual scores, which is again due to guessing. Moreover, this alternative scoring rule biases up the score for all students. Therefore, it is not equivalent, but superior to the correction for guessing formula, since students can respond to all the items randomly without being penalized. Hence, if we offer the students those two alternatives to assess their MCQ exams, they will all clearly prefer the number-right scoring rule.

If we aim to empower students with the option to choose the formula scoring rule in order to reduce the measurement errors introduced by guessing and increase the correlation between knowledge and the score, we should offer them equivalent alternatives. An alternative to the penalty rule aimed at discouraging guessing is a rule that rewards points for unanswered questions, which requires a redefinition of the scoring scale. The aim is to avoid an explicit penalty for incorrect answers while increasing the minimum score to pass the exam. Obviously, this method introduces a kind of penalty—rescaling the scores—but it has the advantage that the students can answer all questions without the pressure of being penalized for incorrect answers; instead, they can be rewarded if they decide not to respond, which could also discourage guessing. Thus, in the experiment we conducted, groups of students were offered the option to choose between the penalty scoring rule (the correction for guessing formula) and the reward scoring rule. By empowering students with the option to choose between equivalent scoring rules, they are assumed and expected to choose the one that best fits their individual characteristics, which could diminish expected bias and variance.[2]

Therefore, this article proposes a flexible tool for correcting MCQ exams, given that students, even those with similar skills and knowledge, are heterogeneous in many other aspects, such as their behavior during an exam, their attitudes toward risk, their guessing behavior, and so on. Moreover, this is interesting precisely because, when a student faces an MCQ exam, her knowledge is not usually binary, but in most cases, she has partial knowledge of the response, that is, she can rule out some options.[3] Under partial knowledge and a given penalty, the expected value of guessing could be greater than the value of omitting. Sometimes, subjects who have partial knowledge omit items with positive expected reward when they are penalized for wrong answers, while they might respond under a reward formula scoring rule. According to Ben-Simon, Budescu, and Nevo (1997), the correction for guessing formula ignores the partial knowledge of students in many cases. Therefore, by providing a method that allows students to choose the scoring rule in an MCQ exam, we contribute to the debate in the literature on the educational measurement of knowledge in relation to the appropriate method of correction (Bar-Hillel, Budescu, & Attali, 2005; Diamond & Evans, 1973; Frary, 1988).

Method

Participants

The target subjects of this study were students attending the Macroeconomics I course for the bachelor of economics degree at the University of Granada during the first semester of 2012–2013. The starting sample contained 189 students.

The subjects were in their natural environment and were unaware of the experiment. In line with the spirit of the well-known classification by Harrison and List (2004), this study represents a natural field experiment, in which subjects have to make simple decisions under risk with nontrivial stakes and potential losses. The experimental observations are consistent with the notion that rational economic behavior occurs only when payoff/losses are tangible or when experiments are nonhypothetical. Students, therefore, have a vested interest in answering the questions to the best of their ability.[4] Thus, individuals choose what they prefer so that any valuation of a chosen alternative reflects real preference. Moreover, this experiment has the advantage that it does not bear costs in terms of money.

Materials

At the beginning of the course, students were given an accurate description of the MCQ and scoring rules (including scales, thresholds, etc.). Therefore, students were previously informed of the following equivalent scoring rules:

Scoring rule 1: The correction for guessing formula, which we call the Penalty Scoring Rule (PSR). The score obtained under PSR (S^0) is given by the following formula:

$$S^0 = R - \frac{W}{M-1},$$

where R represents the number of correct responses, W represents the number of incorrect answers, and M represents the number of alternatives for each question. With no knowledge, the probability of getting a correct answer is $1/M$. However, it should be emphasized that some foils are more appealing than others due to partial knowledge and, therefore, guessing is rarely $1/M$ but higher.

Scoring rule 2: Reward Scoring Rule (RSR). The score obtained under RSR (S^1) is given by

$$S^1 = R + \frac{O}{M},$$

where O is the number of omitted responses. Note that to discourage students from responding at random, they are rewarded for not responding.

In order for both methods to be fully equivalent, the score S^0 and S^1 are considered linearly related such that

$$S^1 = \left(\frac{N}{M}\right) + \left(\frac{M-1}{M}\right)S^0 \qquad (1)$$

where $N = R + W + O$ is the total number of questions, which means that we can make the two scoring rules (PSR and RSR) equivalent by adjusting the score S^1. For instance, suppose a student answers five questions out of 10 ($N = 10$) in an MCQ exam with four alternatives ($M = 4$). If these five questions are correct ($R = 5$) under PSR, her score will be 5. While under RSR, her score would be $5 + 5/4 = 6.25$. In order to make the two methods equivalent, we should require that Equation 1 be fulfilled. Therefore, if the threshold mark to pass the exam using PSR is 5, the threshold mark to pass the exam using RSR will be 6.25. Thus, Equation 1 allows for correspondence between the two scoring rules and can also be seen as the way to rescale the scores with RSR using the same units as PSR in order to make proper comparisons.

In general, students understand PSR very well because they are used to being marked under it. However, they had never been marked under RSR. Therefore, a session was entirely devoted to explaining this "novel" rule to ensure the examiners that the students understood it. In addition, students were provided a table showing the equivalences of both scoring rules.

Notice that regardless of the scoring rule used to correct the exam, students will receive exactly the same mark. A question that could arise with such a statement is the following: What is the difference between using any of the two methods? Scoring rules may affect answering behavior. At the individual level, the number of missing questions (answers) is endogenous and might be different between PSR and RSR. Moreover, most of the students do not know that Equation 1 is fulfilled. Therefore, the students are expected to assume that the scoring rules are simply alternatives. Even though they were aware of Equation 1, they were expected to behave differently when facing either of the two scoring rules. The key issue is that the decision to leave more or fewer items blank will depend on the individual's characteristics, especially attitude toward risk, knowledge level, and the scoring rule. Once the decision is made on all items, the payoff (the score) will be the same regardless of the scoring rule applied.

Experimental Design

In order to avoid the possibility that students were initially grouped according to any specific characteristic, we randomly resort the total enrollment of the official classes A, B, C, and D into four new groups: 1, 2, 3, and 4.[5]

During the semester, four MCQ exams of $N = 10$, which addressed 100% of the syllabus, were conducted for each group of students. Each one lasted about 30 minutes, and they were run at the same time for all groups, each group in a separate classroom.

In each MCQ exam, each group was offered a method of assessment. For two groups, the scoring rules were assigned by the examiner, so that the students did not have the option to choose, as is common practice in MCQ examinations. Therefore, in one of these two groups, PSR was applied, and in the other group, the scoring rule RSR was applied. The other two groups were offered the option to choose the scoring rule. One could choose it through a voting rule. Therefore, students decided the preferred scoring rule (PSR or RSR) to be applied for assessing the MCQ exam through a voting process. Students were allowed to have a quick look at the exam before voting. A simple majority rule was needed. Finally, the last group of students was offered the option of an individual choice. Therefore, each student could freely choose the preferred scoring rule between PSR and RSR.

In sum, the following four methods were allowed:

Method 1: Assigned PSR.

Method 2: Assigned RSR.

Method 3: Choice between PSR and RSR through voting.[6]

Method 4: Individual choice between PSR and RSR.

The experiment was designed so that it did not discriminate against any group, as all groups experienced the four methods. Notice that there were four groups and four MCQ exams during the semester. Therefore, all students had equal opportunities for each of the treatments during the course.

This exam procedure is rather unusual in schools of economics, and most students had probably never had any similar experience.

Hypotheses

Our experimental design allows us to test the following hypotheses:

Hypothesis 1. *PSR and RSR equally discourage guessing, that is, random answers.*

One of the examiners' objectives is to discourage random answers, which is why PSR is widely used. Therefore, if an alternative scoring rule is offered to the students, it should also discourage guessing. Thus, they should be equivalent even across heterogeneous students and expected to discourage guessing equally, and hence, there should be no differences between the distributions of omitted responses across scoring rules.

Hypothesis 2. *There are differences in the scores when choosing the scoring rule, regardless of which rule is chosen, that is, choosing the scoring rule has a value.*

Testing this hypothesis is precisely the main objective of this article. Considering the literature that suggests that the option of choosing has a positive value, intuition suggests that choosing the assessment method should have a positive additional value on the score. Therefore, a student who is provided the option of choosing between PSR and RSR would get a higher score than when not having such an option. However, since this is a novel approach for students, it would not be surprising if unexpected values were obtained.

Hypothesis 3. *There should be no difference in the additional values obtained from choosing PSR or RSR.*

Assuming that Hypothesis 2 is fulfilled, it is expected that the nonzero additional values of choosing PSR and RSR be equal, that is, the difference should not be statistically significant.

Sample Selection Issues and Estimation Strategy

The experimental design provides the proper conditions to know the students' true preferences for PSR or RSR. That is, throughout the course each student i must confront four different tests, τ_j $(j = 1, 2, 3, 4)$, or under the PSR $(\psi = 0)$ or under the RSR $(\psi = 1)$, in which, in each case, the scoring rule can be imposed, that is, no choice $(\rho = 0)$ or choice $(\rho = 1)$. Thus, the expected score values of the individual i in the test $\tau_j, S_{ij}^{\psi\rho}$ can be captured in the diagram of Figure 1.

Due to the design of a continuous system of evaluation used throughout the semester, this research confronted a problem of missing data over time. The source of this bias is based on the fact that students with very low cumulative performances decided not to take the subsequent MCQ exams during the semester. Moreover, other students might have decided, for any other reasons, not to take any of the MCQ exams.

Concisely, even though missing data are common in observational studies due to subjects' self-selection, this situation leads to biased estimates of the true population parameters of linear regression and related models, that is, a difference-in-difference (*DiD*) analysis. Therefore, this research uses a two-stage modeling process by combining a Heckman selection model and a *DiD* analysis that can deal with this selection bias problem, which is regarded as superior to pure cross-sectional difference estimators (Heckman, Smith, & Clements, 1997).

The idea of introducing a Heckman selection method in our *DiD* analysis strategy is rather simple. In the first test, we run a standard *DiD* analysis that includes the population of all subjects ($N_1 = 189$ students). In the subsequent MCQ, we exclude absent students from the data set, leading to a sample population of $N_2 = 145$ in MCQ 2, and so on. This differentiation allows us to establish a prior selection equation of students willing to participate in the subsequent test (Attend = 1). This stage is estimated by a probit regression for all 189 students as follows:

Selection step one: $Pr(Attend = 1|X) = \Phi(X\beta)$, where X refers to the mark achieved in the previous test. Based on this selection regression, the Mills ratio is calculated and included as a covariate in the subsequent *DiD* regression steps to control for the selection bias. The selection procedure is reported test by test in Figure 2.

Ultimately, the effective sample for the third MCQ included 113 subjects and 81 subjects for the fourth MCQ.[7]

We obtain the mean difference of students' scores under choice and no choice both for PSR and RSR to test Hypothesis 2. That is,

$$E\left(S_{ij} \backslash \psi = 0, \rho = 1\right) - E\left(S_{ij} \backslash \psi = 0, \rho = 0\right) = 0$$

$$E\left(S_{ij} \backslash \psi = 1, \rho = 1\right) - E\left(S_{ij} \backslash \psi = 1, \rho = 0\right) = 0$$

DiD $_j$ regression analysis allows us to test the mean difference of students' scores under choice and no choice options for the RSR, and to compare this result with the similar difference obtained for PSR (Hypothesis 3). Thus, to test Hypothesis 3, we ran a *DiD* test strategy according to Card and Krueger (1994) as follows:

$$DiD_j = E\left(S_{ij} \backslash \psi = 1, \rho = 1\right) - E\left(S_{ij} \backslash \psi = 1, \rho = 0\right) - E\left(S_{ij} \backslash \psi = 0, \rho = 1\right) - E\left(S_{ij} \backslash \psi = 0, \rho = 0\right)$$

Results

Figure 3 shows the results of the experiment for the four MCQ exams that were conducted. It shows simple mean scores for the four methods. The results of MCQ 1 are somehow in line with our expectations. However, the results of MCQ 2 are unexpected since the lowest mean scores were obtained under choice. Nevertheless, MCQ 3 and 4 could suggest evidence supporting our expectations.

Figure 4 shows the results for Hypothesis 1, which is not rejected. As can be seen, the distributions of the omitted responses are not statistically different across the PSR and RSR

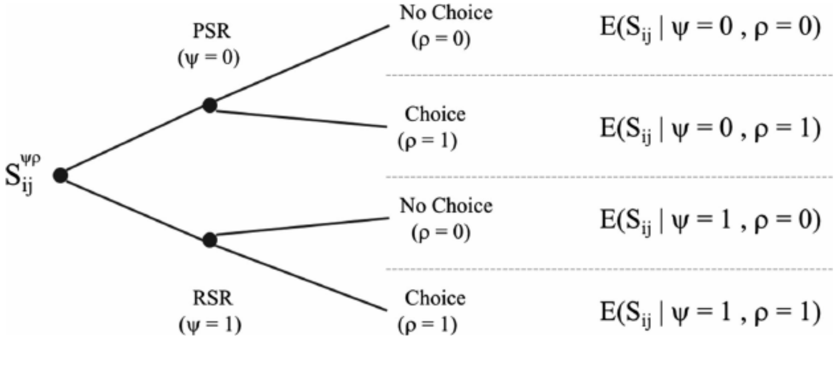

FIGURE 1. Experimental design: Expected score value.

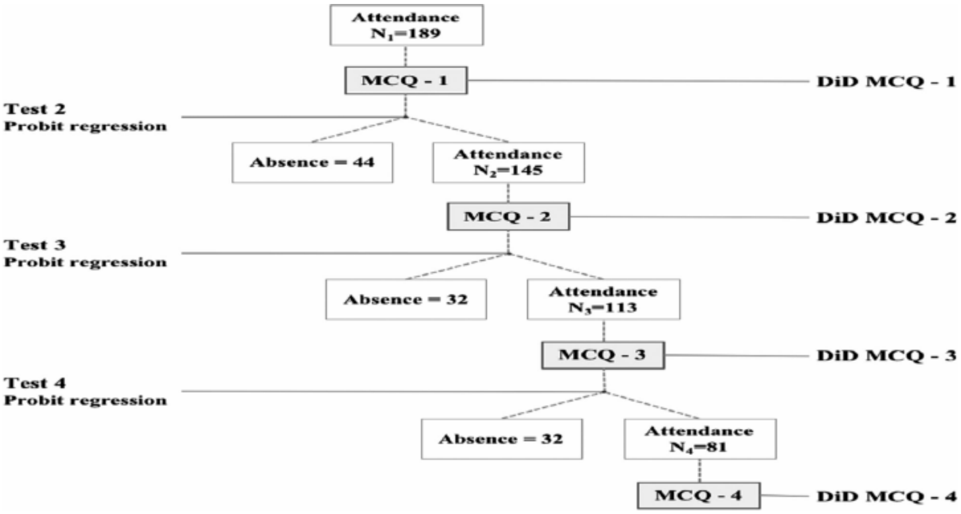

FIGURE 2. Heckman selection strategy.

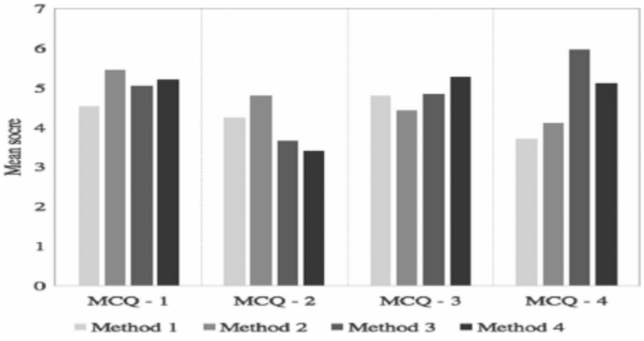

FIGURE 3. Mean scores across MCQ and scoring rule.

for all MCQ exams. These results suggest that there are not enough differences to distinguish the rules based on number of omitted responses and, therefore, they can be offered to the students as alternatives to discourage random answers. Although RSR could be thought to encourage more students with partial knowledge to answer questions, it seems to be offset by the appeal of being rewarded for leaving unanswered questions because this decision implies a sure positive paid instead of a zero paid if the student chose an incorrect answer.[8]

We carried out the *DiD* estimations as described in the previous section, which allows testing Hypotheses 2 and 3. No rejection of Hypothesis 2 would mean that choice has a value. Moreover, no rejection of Hypothesis 3 provides additional

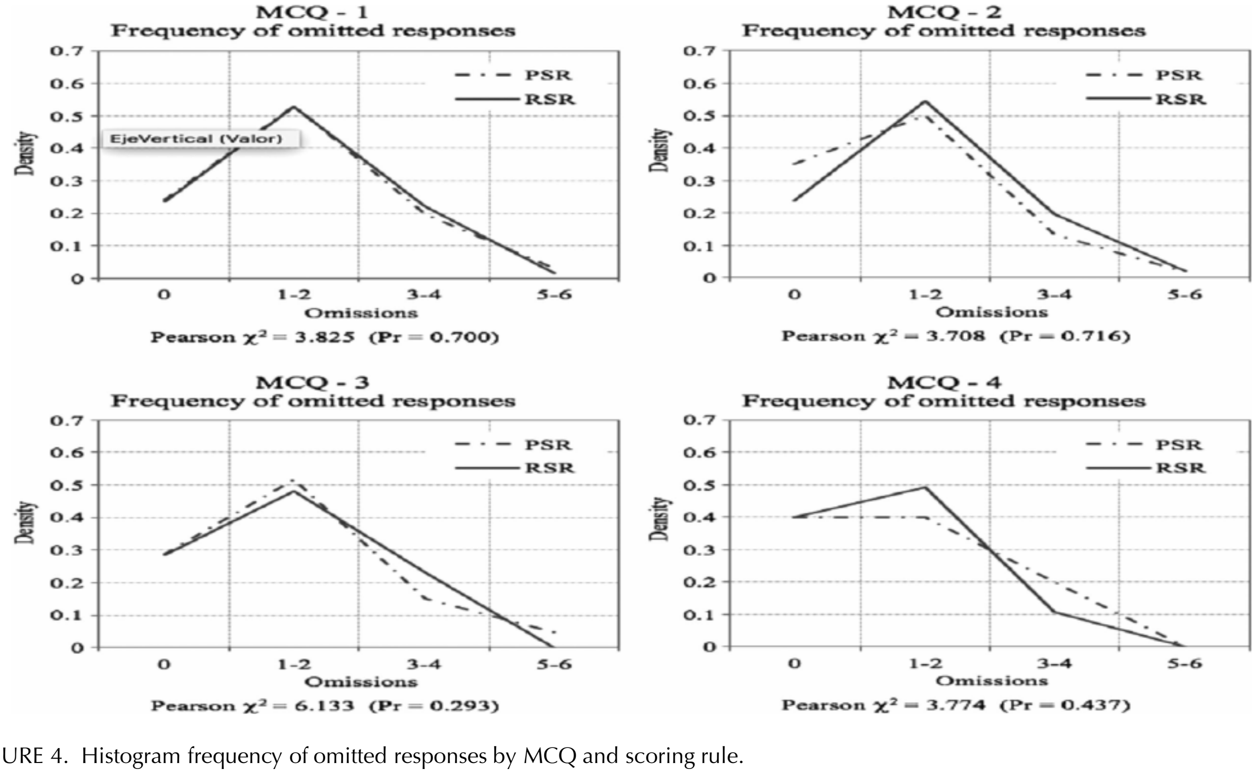

FIGURE 4. Histogram frequency of omitted responses by MCQ and scoring rule.

Table 1. Difference-in-Difference Estimation Multilevel Mixed-Effects Regression by MCQ (Expected Mean Score Controlling by Bias Selection)

	PRS (Base line: $\psi = 0$)			RSR (Follow up: $\psi = 1$)			
	Non choice Control ($\rho = 0$)	Choice Treated ($\rho = 1$)	Difference	Non choice Control ($\rho = 0$)	Choice Treated ($\rho = 1$)	Difference	Difference in Difference
	(a)	(b)	(b)-(a)	(c)	(d)	(d)-(c)	
MCQ - 1	**4.542**	**5.778**	**1.235**	**5.467**	**5.021**	**−0.445**	**−1.681**
Std. error	0.256	0.604	0.656	0.336	0.247	0.417	0.777
t-stat	17.770	9.560	1.880	16.290	20.330	−1.070	−2.160
p-Value	0.000	0.000	0.061*	0.000	0.000	0.287	0.032**
MCQ - 2	**5.800**	**4.984**	**−0.816**	**6.287**	**5.253**	**−1.035**	**−0.218**
Std. error	0.438	0.457	0.486	0.557	0.426	0.476	0.687
t-stat	13.240	10.900	−1.680	11.280	12.340	−2.170	−0.320
p-Value	0.000	0.000	0.095*	0.000	0.000	0.031**	0.751
MCQ - 3	**5.409**	**5.804**	**0.396**	**4.978**	**6.123**	**1.145**	**0.749**
Std. error	0.420	0.360	0.432	0.348	0.482	0.511	0.665
t-stat	12.880	16.120	0.920	14.320	12.720	2.240	1.130
p-Value	0.000	0.000	0.362	0.000	0.000	0.027**	0.263
MCQ - 4	**5.673**	**8.540**	**2.867**	**5.998**	**7.088**	**1.090**	**−1.777**
Std. error	0.688	0.765	0.678	0.764	0.519	0.651	0.930
t-stat	8.240	11.170	4.230	7.850	13.670	1.670	−1.910
p-value	0.000	0.000	0.000***	0.000	0.000	0.098*	0.060*
Multilevel regression	**5.708**	**6.321**	**0.613**	**6.136**	**6.060**	**−0.076**	**−0.689**
Std. error	0.282	0.337	0.306	0.287	0.257	0.237	0.392
t-stat	20.214	18.766	2.003	21.369	23.551	−0.319	−1.757
p-Value	0.000	0.000	0.045**	0.000	0.000	0.749	0.079*
Observations (*N*=528)	132	75		121	200		
	Inverse Mills ratio: −3.7466***			LR test vs. Linear regression: χ^2 = 3.90**			

Note: Means estimated by robust standard errors. Inference: $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

evidence to support the equivalence of these scoring rules. Table 1 shows the results.

Let us concentrate on Hypothesis 2. The fourth column of Table 1 shows the mean difference scores between "no choice" and "choice" PSR. As can be seen, all MCQ exams except MCQ 2 show that choosing has a positive value under PSR. Moreover, in MCQ 1 and 4, the positive values are statistically significant at the 10% and 1% levels, respectively. The negative value of choice found in MCQ 2 is also statistically significant at the 10% level. However, when all MCQ exams are jointly considered, multilevel regression is applied,[9] and we have found a positive and significant value of choice under PSR at the 5% level of significance.

Column 7 of Table 1 shows the mean difference scores between "no choice" and "choice" for RSR. In this case, negative values of choice are found in MCQ 1 and 2, while MCQ 3 and 4 exhibit positive values. However, these values are statistically significant up to the 10% level from the second MCQ on. When all MCQ exams are jointly considered, the multilevel regression shows no statistical evidence of a value of choice.

According to the results, there seems to be a positive value of choice that arises only under PSR. Figure 5 illustrates the results.

Discussion

A likely explanation for the sequential experiment results can be given by a possible learning effect. The evolution of the results over the MCQ exams could suggest that students need experience to learn to choose the scoring rule. It should be stressed that students had never taken an MCQ exam under RSR before. Therefore, MCQ 1 and 2 were useful in allowing the students to become familiar with both scoring rules and to develop useful skills to know how to choose the scoring rule that best fits their individual characteristics. Thus, students in MCQ 3 and 4 who had the option to choose through voting or free election had already been assessed under PSR and RSR in MCQ 1 and 2, so they had previous experience, which could have helped them to make better choices. This is somewhat supported by the results obtained in MCQ 3 and 4 for both scoring rules. In fact, notice that the evidence provided in MCQ 4 could suggest much stronger evidence of the learning-effect argument. Moreover, as can be noticed in Table 1, the mean with PSR under choice steadily increases from test 2 to test 4, while for RSR it increases from test 1. However, this is not the case under no choice for both rules. Therefore, the learning effect is a plausible explanation for the evolution of the results across the tests. Furthermore, the results could also suggest that choice works better for students with higher levels of meta-cognition and self-efficacy/expectancies for success, who are, in turn, believed to be more capable of learning.

Regarding Hypothesis 3, the last column of Table 1 shows that the strongest statistical evidence of the difference in the additional values of choice across scoring rules is found in MCQ 1. In fact, notice that, under PSR, it is positive, while under RSR, it is negative, which could be due to the students' lack of experience in being evaluated under this rule. However, notice that no differences arise in MCQ 2 and 3, and weak evidence is found in MCQ 4 and when all MCQ exams are jointly considered.

Let us provide evidence across gender. In the first panel of Table 2, the second and third columns show the results for males under choice and no choice for PSR, while the fifth and sixth columns show the corresponding results for females. Evidence is provided when all MCQ exams are considered. As can be seen, men and women who chose the scoring rule were better off (columns 4 and 7). However, the

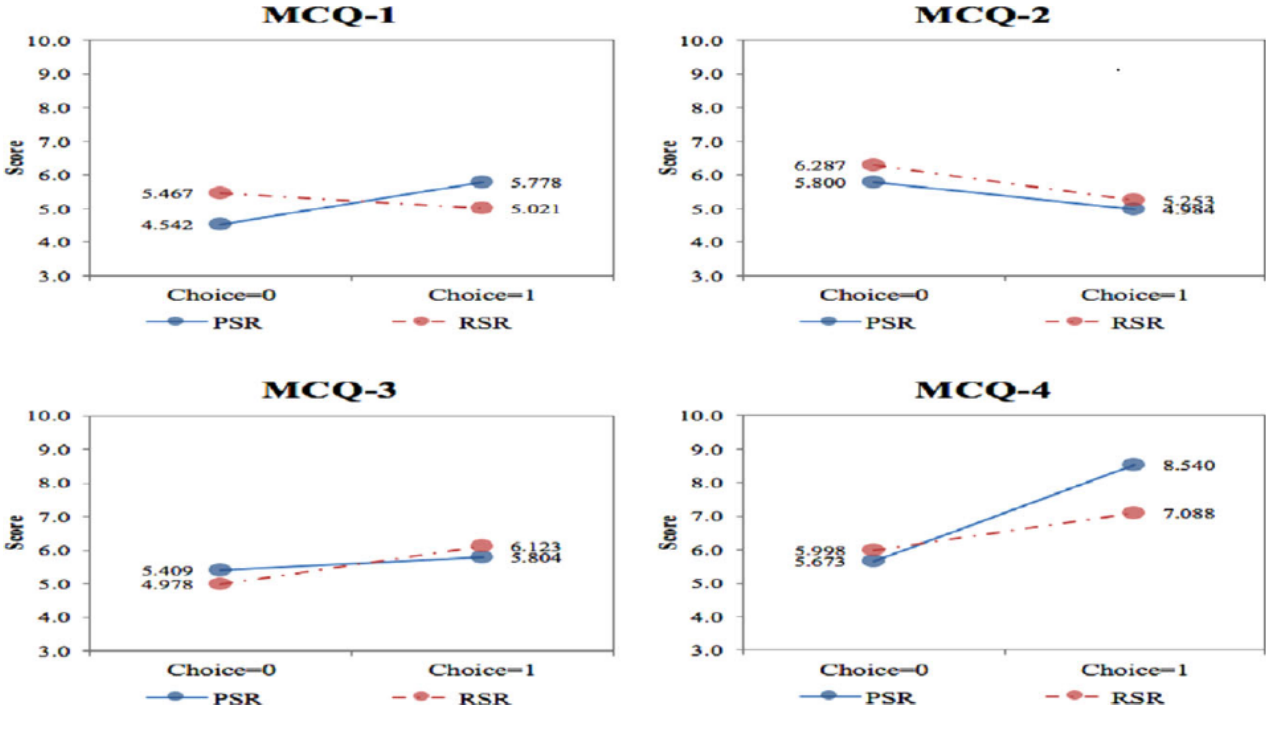

FIGURE 5. Difference-in-difference estimation multilevel mixed-effects regression by MCQ (Expected mean score controlling by bias selection). [Color figure can be viewed at wileyonlinelibrary.com]

Table 2. Difference-in-Difference Estimation by Gender and Type of Scoring Rule Multilevel Mixed-Effects Regression by MCQ (Expected Mean Score Controlling by Bias Selection)

	2.1 Penalty Scoring Rule (PSR)						
	Male (Base line: Sex = 0)			**Female (Follow up: Sex = 1)**			
	Non choice Control (ρ = 0)	**Choice Treated (ρ = 1)**	**Difference**	**Non choice Control (ρ = 0)**	**Choice Treated (ρ = 1)**	**Difference**	**Difference in Difference**
	(a)	**(b)**	**(b)-(a)**	**(c)**	**(d)**	**(d)-(c)**	
Multilevel regression	**5.352**	**5.603**	**0.251**	**5.175**	**6.108**	**0.933**	**0.682**
Std. error	0.313	0.373	0.374	0.330	0.409	0.400	0.547
t-stat	17.093	15.007	0.671	15.680	14.923	2.334	1.247
p-Value	0.000	0.000	0.502	0.000	0.000	0.020**	0.212
Observations (*N*=207)	70	40		62	35		
	Inverse Mills ratio: −2.5584***			LR test vs. Linear regression: $\chi^2 = 1.00$***			
	2.2 Reward Scoring Rule (RSR)						
	Male (Base line: Sex=0)			**Female (Follow up: Sex=1)**			
	Non choice Control ($\rho = 0$)	**Choice Treated ($\rho = 1$)**	**Difference**	**Non choice Control ($\rho = 0$)**	**Choice Treated ($\rho = 1$)**	**Difference**	**Difference in Difference**
	(a)	(b)	(b)-(a)	(c)	(d)	(d)-(c)	
Multilevel regression	**7.357**	**6.479**	**−0.878**	**5.765**	**6.537**	**0.771**	**1.649**
Std. error	0.385	0.322	0.333	0.373	0.346	0.351	0.485
t-stat	19.105	20.148	−2.638	15.442	18.879	2.199	3.400
p-Value	0.000	0.000	0.008***	0.000	0.000	0.028**	0.001***
Observations (*N*=321)	64	103		57	97		
	Inverse Mills ratio: −5.1136***			LR test vs. Linear regression: $\chi^2 = 1.48$			

Note: Means estimated by robust standard errors. Inference: $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

Table 3. Robustness Check Multilevel Mixed-Effects Regression by MCQ (Expected Mean Score Controlling by Bias Selection and Gender Effect)

	3.1 Total Sample						
	PRS (Base line: $\psi = 0$)			**RSR (Follow up: $\psi = 1$)**			
	Non choice Control ($\rho = 0$)	**Choice Treated ($\rho = 1$)**	**Difference**	**Non choice Control ($\rho = 0$)**	**Choice Treated ($\rho = 1$)**	**Difference**	**Difference in Difference**
	(a)	**(b)**	**(b)-(a)**	**(c)**	**(d)**	**(d)-(c)**	
Multilevel regression	**5.845**	**6.454**	**0.608**	**6.274**	**6.204**	**−0.070**	**−0.678**
Std. error	0.290	0.343	0.305	0.295	0.267	0.236	0.391
t-stat	20.125	18.833	1.995	21.260	23.207	−0.295	−1.736
p-Value	0.000	0.000	0.046**	0.000	0.000	0.768	0.083*
Observations (*N*=528)	132	75		121	200		
	Inverse Mills ratio: −3.7112***			LR test vs. Linear regression: $\chi^2 = 3730$**			
	3.2 Considering only Voting Choice						
	PRS (Base line: $\psi = 0$)			**RSR (Follow up: $\psi = 1$)**			
	Non choice Control ($\rho = 0$)	**Voting choice Treated ($\rho = 1$)**	**Difference**	**Non choice Control ($\rho = 0$)**	**Voting choice Treated ($\rho = 1$)**	**Difference**	**Difference in Difference**
	(a)	(b)	(b)-(a)	(c)	(d)	(d)-(c)	
Multilevel regression	**5.860**	**6.457**	**0.597**	**6.308**	**6.345**	**0.036**	**−0.561**
Std. error	0.278	0.455	0.436	0.283	0.288	0.269	0.515
t-stat	21.075	14.192	1.371	22.316	22.014	0.135	−1.090
p-Value	0.000	0.000	0.170	0.000	0.000	0.893	0.276
Observations (*N*=396)	132	28		121	115		
	Inverse Mills ratio: −3.4852***			LR test vs. Linear regression: $\chi^2 = 0.04$			
	3.3 Considering only Free Choice						
	PRS (Base line: $\psi = 0$)			**RSR (Follow up: $\psi = 1$)**			
	Non choice Control ($\rho = 0$)	**Individual choice Treated ($\rho = 1$)**	**Difference**	**Non choice Control ($\rho = 0$)**	**Individual choice Treated ($\rho = 1$)**	**Difference**	**Difference in Difference**
	(a)	(b)	(b)-(a)	(c)	(d)	(d)-(c)	
Multilevel regression	**5.880**	**6.490**	**0.610**	**6.324**	**6.104**	**−0.220**	**−0.830**
Std. Error	0.275	0.367	0.337	0.279	0.293	0.281	0.440
t-stat	21.396	17.687	1.808	22.650	20.798	−0.783	−1.888
p-value	0.000	0.000	0.071**	0.000	0.000	0.434	0.059*
Observations (N=385)	132	47		121	85		
	Inverse Mills ratio: −3.8066***			LR test vs. Linear regression: $\chi^2 = 0.3434$			

Note: Means estimated by robust standard errors. Inference: $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

difference between choice and no choice is significant only for women. Moreover, no difference is found between the additional value obtained for men and women who choose PSR (column 8).

The second panel of Table 2 shows a similar test for RSR. When all MCQ exams are jointly considered, we found that men who chose RSR obtained a negative significant value. On the contrary, women who chose RSR obtained an additional positive significant value. Therefore, the *DiD* estimation shows evidence suggesting differences between men and women when choosing RSR. This result could explain the results obtained in Table 1, in which no evidence was found between choosing and not choosing RSR since those opposite values for choice for males and females seem to cancel each other out.

As suggested by Espinosa and Gardeazabal (2010, 2013) and Prieto and Delgado (1999), there can be differences across gender under an MCQ exam provided it involves risky decisions. Moreover, the psychological and experimental economic literatures have been very active in providing evidence on gender differences in risk. The meta-analysis by Byrnes, Miller, and Schafer (1999) found greater risk taking in males and Charness and Gneezy (2011) found evidence supporting their results. Therefore, the option of choosing the scoring rule could be helpful in providing a fairer method for students under MCQ exams. The results in Table 2 show that under no choice, males got on average higher scores, which is in line with the literature suggesting that males perform better in MCQ exams (Beller & Gafni, 2000; Ben-Shakhar & Sinai, 1991; Bolger & Kellaghan, 1990; Prieto & Delgado, 1999). However, it can be also noticed in Table 2 that providing the option to choose the scoring rule made females get, on average, higher scores.

In addition, the first panel of Table 3 shows multilevel fixed effect regressions controlling for gender effects. As can be noticed, the results hardly change with respect to Table 1.

Moreover, remember that we offered the students two ways of choosing: the voting approach (method 3) and the individual free choice (method 4). Therefore, we want to know if there is any difference between such methods. The second panel of Table 3 shows the results when students choose the scoring rule through voting compared with the no choice option under both PSR and RSR. As can be seen, although the option of choosing the scoring rule through voting adds a positive additional value to the score, the differences are not statistically significant, which suggests that choosing through an “electoral process” does not make students better off. However, as can be seen in the third panel of Table 3, the previous results obtained in Table 1 hold. Therefore, positive value arises from the option of choosing the scoring rule when students freely choose it and under PSR.

The results provided in this article may support the idea that the mere exercise of choice can provide a sense of autonomy, control, and empowerment with positive consequences for some students.

Conclusions

This article contributes to the literature on the value of choice and to the literature on education, specifically regarding the topic of the educational measurement of knowledge. We propose a novel experimental methodology in which groups of students were offered the option to choose between two equivalent scoring rules to assess MCQ. Students were offered the traditional correcting for guessing formula, which we call PSR. Alternatively, a scoring rule that rewards points for omitted responses was also offered (RSR). We randomly split the total sample into four groups of students. Two groups of students had the option to choose between PSR and RSR. However, the other two groups of students did not, and one was assessed using PSR and the other using RSR. We are interested in quantifying the value of choice, and we assume that students chose the scoring rule that best fit their individual characteristics. Therefore, we compare the groups that chose either PSR or RSR with the corresponding groups that were not allowed to choose. The main results show *(i)* that it is possible to offer the students two statistically equivalent scoring rules for discouraging guessing, that is, random answers; and *(ii)* in general, there seems to be evidence in favor of a positive value of the option of choosing the scoring rule to assess an MCQ exam. Moreover, the sequential empirical test suggests that students need to learn to take advantage of the option of choosing the scoring rule. Additionally, *(iii)* the results show that women obtain greater benefits from the option of choosing the scoring rule.

Notes

[1] Philosophy and neural science have also studied the value of choice.

[2] While the two scoring rules are mathematically equivalent, they may not be psychologically equivalent.

[3] Misinformation is another issue that should be taken into account. Sometimes the answer is wrong not due to guessing incorrectly, but to a fundamental mistake when the examinee actually believes that her answer is correct.

[4] A common critique of the experiments is that, on the one hand, if they are hypothetical, subjects could take them lightly. On the other hand, if they are nonhypothetical and money-based, they often become too expensive for an instructor to conduct.

[5] Some demographic information for the sample is shown in Appendix 1.

[6] Students were informed which scoring rule had been decided just before taking the test. Therefore, students that vote for the losing option are expected to behave similarly to when the rule is imposed. They could have thought that they would have performed better with the other rule but they had to adapt to the winning rule.

[7] The high abandonment rate during the continuous assessment might have been due to the fact that students had the alternative of a final exam after finishing the course.

[8] The results hold under choice and no choice, and across gender. Available upon request.

[9] This approach results in a statistical model where parameters are allowed to vary at more than one level. Considering this method, the effect of differences in the difficulty of each test is isolated. Details on this estimation method are provided in Appendix 2.

References

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*(1), 31–36.

Bar-Hillel, M., Budescu, D., & Attali, Y. (2005). Scoring and keying multiple choice tests: A case study in irrationality. *Mind and Society*, *4*(1), 3–12.

Beller, M., & Gafni, N. (2000). Can item format (multiple choice vs. open-ended) account for gender differences in mathematics achievement? *Sex Roles*, *42*(1–2), 1–21.

Ben-Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement*, *28*, 23–35.

Ben-Simon, A., Budescu, D. V., & Nevo, B. (1997). A comparative study of measure of partial knowledge in multiple-choice tests. *Applied Psychological Measurement*, *21*(1), 65–88.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement*, *27*, 165–174.

Bredon, G. (2003). Take home tests in economics. *Economic Analysis and Policy*, *33*(1), 52–60.

Byrnes, J., Miller, D. C., & Schafer, W. D. (1999). Gender differences in risk taking: A meta-analysis. *Psychological Bulletin*, *125*, 367–383.

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, *84*, 772–93.

Charness, G., & Gneezy, U. (2011). Strong evidence for gender differences in risk taking. *Journal of Economic Behavior and Organization*, *83*(1), 50–58.

Diamond, J., & Evans, W. (1973). The correction for guessing. *Review of Educational Research*, *43*(2), 181–191.

Dowding, K., & John, P. (2009). The value of choice in public policy. *Public Administration*, *87*(2), 219–233.

Espinosa, M. P., & Gardeazabal, J. (2010). Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology*, *54*, 415-425.

Espinosa, M. P., & Gardeazabal, J. (2013). Do students behave rationally? Evidence from a field experiment. *Journal of Economics and Management*, *9*(2), 107–135.

Frary, R. B. (1988). Formula scoring of multiple-choice tests (correction for guessing). *Educational Measurement: Issues and Practice*, *7*(2), 33–38.

Glimcher, P. W. (2004). *Decisions, uncertainty, and the brain: The science of neuroeconomics*. Cambridge, MA: MIT Press.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). Oxford, UK: Oxford University Press.

Greene, W. H. (2002) *Econometric analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature*, *42*, 1009–1055.

Heckman, J. J., Smith, J., & Clements, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *Review of Economic Studies*, *64*, 487–535.

Iyengar, S. S., & Lepper, M. R. (1999). Rethinking the value of choice: A cultural perspective on intrinsic motivation. *Journal of Personality and Social Psychology*, *76*, 349–366.

Luke, D. A. (2004). *Multilevel modeling*. Newbury Park, CA: Sage.

Prieto, G., & Delgado, A. R. (1999). The role of instructions in the variability of sex-related differences in multiple-choice tests. *Personality and Individual Differences*, *27*, 1067–1077.

Siegfried, J. J., Saunders, P., Stinar, E., & Zhang, H. (1996). Teaching tools: How is introductory economics taught in America? *Economic Inquiry*, *34*(1), 182–192.

Srholec, M. (2010). A multilevel approach to geography of innovation. *Regional Studies*, *44*, 1207–1220.

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2004). Matching behavior and the representation of value in the parietal cortex. *Science*, *304*(5678), 1782–1787.

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2005). Choosing the greater of two goods: Neural currencies for valuation and decision making. *Nature Reviews Neuroscience*, *6*, 363–375.

Usher, M., Elhalal, A., & McClelland, J. L. (2008). The neurodynamics of choice, value-based decisions, and preference reversal. In N. Chater and M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (277–300). Oxford, UK: Oxford University Press.

Appendix 1

Demographic Information

Students were asked to report some information about income, parents' educational level, and so on. As can be noticed, the sample is fairly homogeneous.

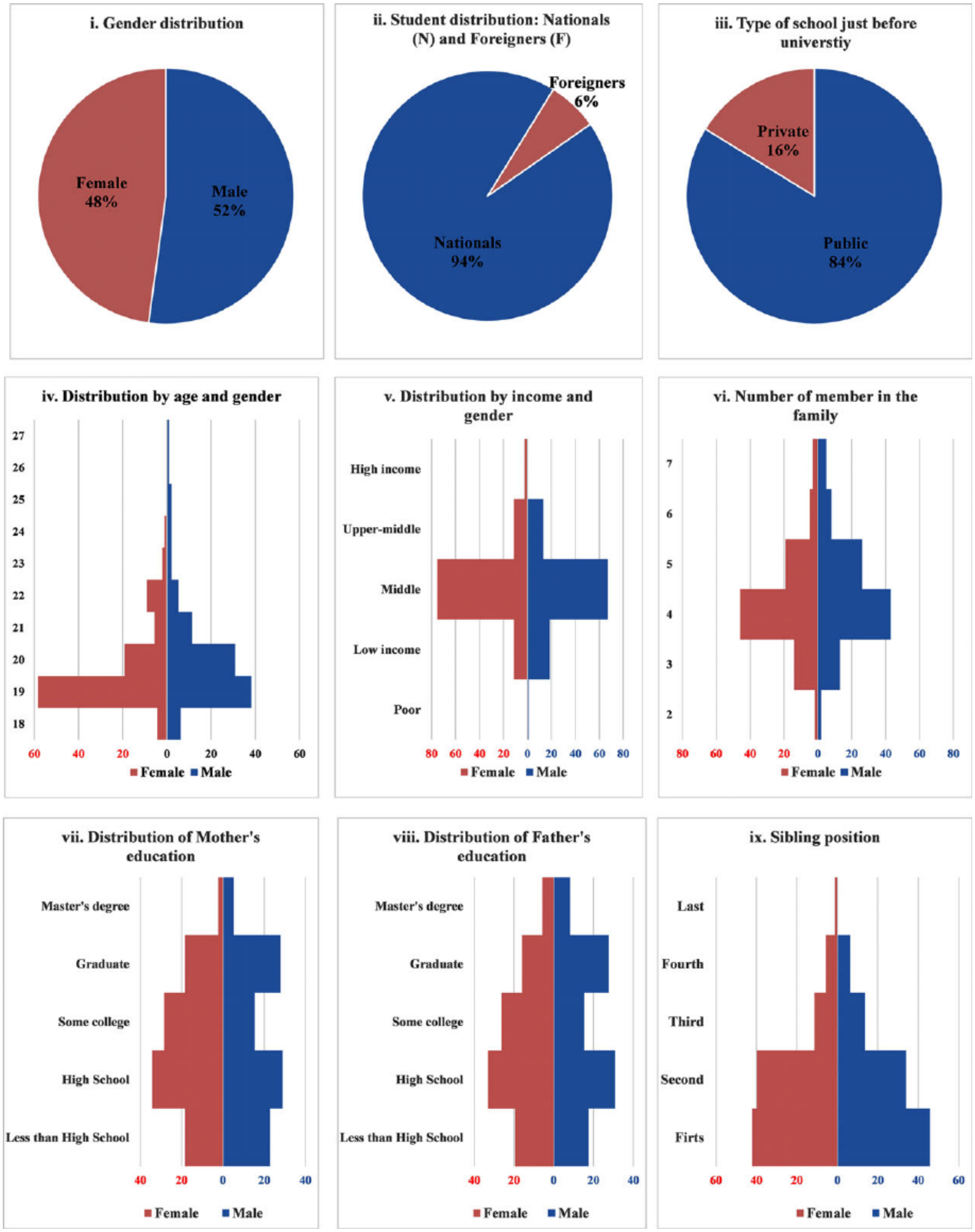

FIGURE A1. Some demographic information for the sample. [Color figure can be viewed at wileyonlinelibrary.com]

Appendix 2

Multilevel Regression

A multilevel mixed-effects regression was run to estimate the expected mean score value under choice and no choice both for PSR and RSR and obtain the mean difference in student's scores. In a multilevel mixed-effect analysis, sometimes also called a hierarchical, random coefficient or nested data model, the data structure in the sample population is hierarchical, and data are viewed as a multistage sample from this hierarchical population (Goldstein, 2003). Consequently, students are hierarchically nested in a two-level model that relates the dependent variable to predictor variables at more than one level (Luke, 2004). First, the macrolevel contains the four different tests (τ_j), and second, i students are assumed to be randomly sampled per unit (microlevel). Formally, we can write a generalized linear two-level model for a generic student of any of groups as

$$E\left(S_{ij}^{\psi\rho}\right) = \beta_0 + \beta_1 SR_{ij} + \beta_2 CH_{ij} + \beta_3\left(SR_{ij} \times CH_{ij}\right) + E\left(\varepsilon_{ij}\right), \tag{A.1}$$

where $E\,(S_{ij}^{\psi\rho})$ is the expected value of score of student i in test j, SR_{ij} is a dummy variable that takes the value of 1 if student i was marked under RSR in test j and takes the value of zero under PSR, and CH_{ij} is a dummy variable that takes the value of 1 if student i had the option to choose the scoring rule in test j and takes the value of zero if the student did not have such an option. β_0, β_1, β_2, and β_3 are the parameters to be estimated.

Finally, ε_{ij} is an error term that, in the hierarchical model, consists of two components:

$$\varepsilon_{ij} = \gamma_i + \mu_{ij},$$

where ε_{ij} is the individual-specific random effect, that is, it is the deviation of the ith student's score from the average for the jth test, γ_i is the specific individual effect, and μ_{ij} is an independent and identically distributed (iid) disturbance with $E\,(\mu_{ij}) = 0$. As noted by Srholec (2010), the presence of more than one residual term makes standard multivariate models, such as fixed-effects specification, inapplicable, and generalized maximum likelihood (GML) procedures should therefore be used to estimate these models properly.

By inspecting the Equation A.1, we should be able to notice that the coefficients have the following interpretation: β_0 is the constant term in the regression, β_1 is the expected added value of RSR (to account for average permanent differences between PSR and RSR), β_2 is the expected added value that provides the option of choosing the scoring rule and is common to PSR and RSR, and β_3 is the true mean difference of students' scores considering the scoring rule and the option of choosing it.

We follow the generalized Heckman approach as developed by Greene (2002) to compute the inverse Mills ratio (λ_{ij}), and the selection bias was corrected by including this Mills ratio when Equation A.1 was estimated. Thus, the conditional expectations of the score values, conditional on taking four different tests, can be written as a single equation:

$$E\left(S_{ij}^{\psi\rho}\right) = \beta_0 + \beta_1 SR_{ij} + \beta_2 CH_{ij} + \beta_3\left(SR_{ij} \times CH_{ij}\right) + \varphi\lambda_{ij}\sigma_\varepsilon + \gamma_i,$$

where φ is the correlation between the unobserved determinants of taking different tests designed with different levels of difficulty and the unobserved error term ε_{ij} and σ_ε is the *SD* of ε_{ij}.