Gender Bias in Test Evaluation

Carmen Marín Alfonso Rosa-García

Abstract

We explore the differences in the output in test evaluation across a sample of 2372 exams done by undergraduate students from the Universidad de Murcia from 2005 until 2010. We find a gender bias in the amount of answered questions that can be associated with gender differences in risk aversion. The gender gap in scores is weak. We find also distorting effects of test scrambling on some of the students. Our results suggest that a good design of test exams should prevent risk effects.

Keywords: test evaluation, gender differences

JEL Classification: A22, I21, J16
1 Introduction

We analyze 8 exams completed by undergraduate students from 2005 until 2010 in the subject Political Economy from Degree in Business in the Universidad de Murcia. This subject was taught in the second semester of the first year of the degree. In our sample, we have a total of 2372 exams in which we have explored the existence of gender difference in test exam behavior. Our results show that, consistently across all the exams, males tend to answer a higher amount of questions than females, which we argue is in line with the empirical and experimental evidence on the higher risk aversion observed in women. Our data show a mixed evidence on the existence of a bias in score output, mainly in favor of males but not always. This observation is also consistent with the previous literature: from the '70s, it has been observed a gender bias in test output in Economics, in whose test exams females tend to get lower marks than males, but only slightly. We also explore the effect of test scrambling. We do not find consistent and general effects of test scrambling. This evidence is also in line with the previous literature. However, test scrambling seems to distort the results for some subgroups of students, although in different directions. These distortions support the deletion of a scrambling and non-scrambling test.

Most of these effects are consistent with the high and low mark subgroups of agents. Therefore our data are consistent with previous evidence and we add two new insights: a link between test behavior and the higher risk aversion of females, as well as a partial effect of test scrambling.

The gender bias in test score in Economics has been deeply studied for years (Siegfried, 1979; Ballard and Johnson, 2005). These works usually show a higher score for males, but not always. These results are in line with the gender gap in favor of men usually reported in math scores (Hedges and Nowell, 1995). Interestingly, this bias seems to be correlated with women culture acceptance, and seems to decrease across years (Guiso et al., 2008). In our sample, we find that males improve scores of girls in 5 of the 8 exams, and difference was significant only in 2 cases. This evidence supports the weakness of the gender bias in favor of males in this kind of exams.

The main contribution of our work is the evidence of a higher degree of risk in males answer, measured as the amount of question answered. This effect is robust significant in all the exams of our sample, and show that females tend to answer around 5% less questions than males. In the last years, it has been shown an increasing amount of evidence that support a higher risk aversion of women (Croson and Gneezy, 2009). The

1 Makridou-Bossion (2003) show evidence of the opposite effect for Greek students. In that case females behaves slightly better than men in test exams. They relate this difference with the characteristic of Greek education system, which tend to emphasize more memory than analytical aptitudes.

2 Although differences in mean seem to be correlated with cultural factors, variability seems to be worldwide larger in the case of males (Machin and Pekkarinen, 2008).
exams in our sample are designed with a penalty for getting an expected 0 score for a random respondent. However, this method widely used penalized those agents who are more risk averse (Burgos, 2005).

According to the higher amount of questions answered by males, we find that they significantly obtain a higher level of correct answers in 4 of the exams, and obtain a higher level of incorrect answers in 2. The addition of these two effects, where the higher number of correct questions tend to dominate, generates that males obtained a significantly higher mark in the 3 of our 8 exams. Since risk aversion makes that agents answer less questions from optimal (with the usual penalization), this evidence suggests that risk aversion can be at least a partial source of gender bias in test scores.

As well as the effects of gender in test performance, the literature has also explored the effects of scrambling test. Typically it has been found non significant differences (or very slight) between those students who solve the exam with questions in a randomized order against those students who solve in according with the sequence of topic teached in the class (Gohmann and Spector, 1989). Our sample support this point of view off the general sample. However, we find sistematic differences between some groups of agents when they solve the randomized or the ordered version of the exam. Usually (but not always), these differences lead to increase their answers to the males with low level in our sample, when solving the non-ordered versions of the exam. This supports the necessity of presenting the exam with the same random generation for all agents in order to prevent undesired distortions.

Since our data suggest that risk aversion may be the source of the gender bias observed in economic test exams, measures tending to decrease the risk in this kind of exams may help to decrease this bias.

2 The Data

We have analyzed 8 exams of Political Economy across a period of 6 years. We analyze all the exams where we were able to recover all the marks, which includes the June and September exams in 2005, 2008 and 2009, the exam in September 2007 and in June 2010. Table 1 reports some descriptive statistics of the exams. We have a total of 2372 exams, from a minimum of 161 in September 2005 and a maximum of 485 in June 2005.
The subject Political Economy was taught in the second semester of the first year of the 3-years degree in Business in the Universidad de Murcia. This was a typical introductory course to Microeconomics. The subject was taught until the course 2008-09, when the old 3-year Degree was transformed according to the Bolonia program. During its usual period, the final exam included a first 20-question test part, which accounted for the 40% of the final mark and that the student had to pass in order to be evaluated in the subject. We label these exams as type A, and we include also in our sample the first of the exams done after the subject finished. The students have still 2 years in order to pass the subject, but now all the exam is a test type with 40 questions. We have labelled the exam we include in our sample (June 2010) as type B.

From the total of 2372 exams, we have 1244 done by females (52.45%). Score and answers are slightly higher in males than in females.

We find that differences between males and females in answered questions are significant in each of the 8 exams. Differences between sexes in the score obtained are significant in the exam of 2008, but not in the rest of exams. This difference is generated by a significant difference in the amount of correct questions in that exams. In the exams of September 2007 and June 2010 we also find a significant difference in the amount of correct answers by sex, but it is compensated by a significant difference in the amount of incorrect answers (also higher for males). Incorrect answers are also significantly higher for males in September of 2008 and June of 2009, but they do not generate a significant difference in marks.

Therefore, when there exists a significant gender bias in score it is generated by a higher proportion of correct answers by males, who typically are riskier in their behavior. Therefore, when this lowest risk aversion generates a higher proportion of correct answers, males tend to obtain higher scores.
2.1 Gender, test scrambling and levels

2.1.1 Gender and levels

In order to analyze how students of different sex or levels confront the exams, we have estimated the effect of gender as explanatory variables of score, correct, incorrect and non-answered questions, and later we study it combined with the effect of levels. Table 2 reports the significance of gender for each regression in each of the exams:

<table>
<thead>
<tr>
<th></th>
<th>J05</th>
<th>S05</th>
<th>S07</th>
<th>J08</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>-0.01 (0.07)</td>
<td>0.05 (0.12)</td>
<td>-0.18 (0.11)</td>
<td>-0.14* (0.08)</td>
</tr>
<tr>
<td>Correct</td>
<td>-0.23 (0.29)</td>
<td>-0.07 (0.44)</td>
<td>-0.87** (0.40)</td>
<td>-0.81*** (0.30)</td>
</tr>
<tr>
<td>Incorrect</td>
<td>-0.33 (0.22)</td>
<td>-0.49 (0.39)</td>
<td>0.02 (0.04)</td>
<td>-0.26 (0.28)</td>
</tr>
<tr>
<td>Unanswered</td>
<td>0.56*** (0.18)</td>
<td>0.56* (0.32)</td>
<td>0.86*** (0.30)</td>
<td>1.07*** (0.23)</td>
</tr>
<tr>
<td></td>
<td>S08</td>
<td>J09</td>
<td>S09</td>
<td>J10</td>
</tr>
<tr>
<td>Score</td>
<td>0.09 (0.10)</td>
<td>0.02 (0.09)</td>
<td>-0.04 (0.10)</td>
<td>-0.26 (0.18)</td>
</tr>
<tr>
<td>Correct</td>
<td>0.04 (0.35)</td>
<td>-0.18 (0.32)</td>
<td>-0.38 (0.39)</td>
<td>-0.95* (0.57)</td>
</tr>
<tr>
<td>Incorrect</td>
<td>-1.01*** (0.38)</td>
<td>-0.57* (0.33)</td>
<td>-0.40 (0.35)</td>
<td>0.17 (0.47)</td>
</tr>
<tr>
<td>Unanswered</td>
<td>0.98*** (0.27)</td>
<td>0.86*** (0.30)</td>
<td>0.78*** (0.26)</td>
<td>0.78* (0.47)</td>
</tr>
</tbody>
</table>

Table 2. Coefficient of gender (0 for males, 1 for females) in the regression of the dependent variables in first column.

Standard deviations in parenthesis. Coefficients are significant at 1% (***) or 10% (*).

The effect of gender on scores is not clear. In some exams females obtained higher marks, although in 5 of the 8 exams were males who obtained. In any case, gender is significant only in one exam (in which males got higher scores). However, gender is always significant when testing about answered questions, with females answering always less questions. It is reasonable to assume that this is the result of different risk aversions, given the stylized fact of women being more risk averse than men (Gneezy and Croson, 2009). The very small differences in scores support this view.

We find therefore a significant and consistent gender bias in the amount of unanswered questions. In order to determine if this occurs to the more or less prepared students, we estimate the effect of gender controlling by if the student was in the upper or in the lower part of the score distribution. Table 3 shows the results:
Table 3. Effect of gender (0 for males, 1 for females) and level (0 for the lower and 1 for the upper) on the amount of unanswered questions. Coefficients are significant at 1% (***), 5% (**) or 10% (*).

In all the exams, we find that the interaction effect between gender and level is not significant, meaning that gender and level differences are independent. The effect of gender was significant in this regression in 5 of the 8 exams, including the 3 exams of June before the exam format was changed in 2010. One would expect that level were always significant, since usually those students with higher marks are supposed to answer more questions. It is not always the case, supporting the idea of unanswered questions being also a consequence of something further knowledge. Risk aversion is a reasonable explanation, which agrees with our view of (at least partially) causing the behavior of females.

2.1.2 Test scrambling and levels

We want to explore also if test scrambling can be a source of differences among students. When doing the exams, we provide the students with four different types, the first of which was ordered according to the order followed in the lessons in class, while the other 3 types were randomly ordered. Significantly, we do not find any significant effect of scrambling test neither in the amount of unanswered questions nor in the score, correct or incorrect questions in any of the 8 exams.

When controlling by the level of the student (if the student was in the upper or in the lower part of the score distribution), the effect of test scrambling is once more always non significant, and the interaction between test scrambling and level was significant only for June of 2008.

Therefore we do not find a significant effect of test scrambling in any of the 8 exams. However, the estimated coefficient was always positive, when controlling and not controlling by the level. This means that in all the exams people with the exam presented in an ordered manner answered less questions, although
the effect was never significant. However, the repetition of the sign in all the cases suggests the possible existence of a slight effect of test scrambling in the behavior of students.

### 2.1.3 Gender, test scrambling and levels

Finally, we have combined gender and test scrambling in order to explore the existence of, at least, some effects over different subjects. In the complete regression, we have tested the effect of gender and scrambling controlling by the various levels in scores in our sample, separating our data in quartiles. The equation we estimate is

\[ DV_i = \beta_0 + \beta_1 Sex_i + \beta_2 ST_i + \beta_3 q1_i + \beta_4 q2_i + \beta_5 q3_i + \beta InteractionEffects_i + \varepsilon_i \]

where the Dependent Variable \( DV \) is the score, or amount of correct, incorrect and nonanswered questions, \( Sex \) is a dummy variable with the value 0 for males and 1 for females; \( ST \) takes the value 1 if the subject did the exam with the questions ordered as taught in class and 0 if the questions were ordered at random; \( qj \) is a dummy variable that takes the value 1 if the subject was in the \( j \)–th quartile by score; and \( InteractionEffects \) is a set of variables that take into account the interaction effects among the explanatory variables.

We do not find any consistent effect (i.e., significant in all the exams) further the lower level of answered questions by females. Table 4 synthesizes the most relevant findings

### Table 4. Facts on the interaction between type of exam and gender that are significant in more than 1 of the exams.

<table>
<thead>
<tr>
<th>Fact: Unanswered questions</th>
<th>#(Significant exams)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Males with ordered exam answer less questions&quot;</td>
<td>2/8</td>
</tr>
<tr>
<td>&quot;Females with ordered exam get more correct questions&quot;</td>
<td>3/8</td>
</tr>
<tr>
<td>&quot;Males with ordered exam get less incorrect questions&quot;</td>
<td>2/8</td>
</tr>
<tr>
<td>&quot;Males with ordered exam get less score&quot;</td>
<td>2/8</td>
</tr>
<tr>
<td>&quot;Low level males with ordered exam answer less questions&quot;</td>
<td>4/8</td>
</tr>
<tr>
<td>&quot;Low level males with ordered exam get less incorrect questions&quot;</td>
<td>3/8</td>
</tr>
<tr>
<td>&quot;High level with ordered exam answer more questions&quot;</td>
<td>2/8</td>
</tr>
<tr>
<td>&quot;Low level with ordered exam answer less questions&quot;</td>
<td>2/8</td>
</tr>
</tbody>
</table>

When studying the interaction between scrambling tests, gender and levels, we find effects of the ordered exam in some of the groups of students. In particular, our sample suggests that having the exam ordered

---

\(^3\)All the regressions are available upon request.
have effects of different sign depending on gender and level, which may explain the non general significance of test scrambling. In particular, when subjects do the exam in an ordered way, this generates sometimes a decrease in the amount of answered questions by high level subjects with an increase of answered questions for low level subjects: low level subjects tend to be more risk averse and high level subjects less risk averse when having the questions in an ordered way. This effect tend to affect specially the males of low level, which become more risk averse and make them to do less mistakes in a significant amount of exams. Moreover, it is also observed an effect of the ordered exam in the amount of correct questions answered by women.

In general, we observe few effects of the exam in an ordered way, but when it has an effect usually it improves the behavior of the agents. Exams in the ordered way seem to mitigate the possible perverse effects of risk aversion.

3 Conclusion

In our data on test exams, males and females differ mainly in the level of non answered questions, which can be easily explained as a result of the consistently higher risk aversion of females. This different level of risk aversion seems to be an important reason for the gender bias observed in scores, although penalization tend to reduce it. However, a lower degree of risk aversion can generate a higher expected mark and following this reasoning, is likely that the slight male advantage in the test exam score hides a subyacent higher knowledge by females which are penalized because of their higher risk aversion.

We do not find general effects of test scrambling, although it can be identified different behavior of some groups when solving an ordered or a randomized test exam. This suggest that test scrambling really affects students behavior, although in an unclear way.

In order to solve these problems, scoring rules that reduce the possibility of different risk attitudes could be implemented. For instance, it could be used a version where agents should answer all questions compulsory, still having a penalty for wrong questions. This should be fair since expected value remains constant, although it could generate self-selection problems: the higher risk of this exams could generate that people who is more risk averse tend to not go to the exams in a higher proportion.
4 References

References


