Economics of Education Review 59 (2017) 43–62

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Shying away from demanding tasks? Experimental evidence on gender differences in answering multiple-choice questions

Gerhard Riener, Valentin Wagner*

Düsseldorf Institute for Competition Economics, Universitätsstraße 1, 40225 Düsseldorf, Germany

ARTICLE INFO

Article history:
Received 13 December 2016
Revised 28 June 2017
Accepted 28 June 2017
Available online 29 June 2017

JEL classification:
I20
C93
D80

Keywords:
Educational economics
Behavioral decision-making
Answering strategies
Gender differences
Field experiment
Multiple-choice test

ABSTRACT

Access to higher education is one important prerequisite for later employment possibilities. Often access is regulated *inter alia* by multiple-choice entrance exams. The application of this testing format is problematic if it favors the answering strategies of certain groups. We present the results of a field experiment in answering multiple-choice questions. Our sample consists of 2113 pupils from different school types. We find that girls skip more answers than boys only if the questions are difficult. This gender gap vanishes when extrinsic rewards are provided. This suggests that our findings are compatible with a stereotype threat explanation. Moreover, the gender gap is found only for pupils of school types preparing for the academic track. It is therefore important to consider all social strata in the design of testing formats.

1. Introduction

In recent years there has been an increase in the inequality of wages across groups within many societies, which is driven by the returns to formal education (Lemieux, 2006). Moreover, there is an ever increasing wage gap between socially disadvantaged groups (Autor, Katz, & Kearney, 2008) and although the gender gap in educational achievement is shrinking or has even been reversed in most subjects (see Niederle & Vesterlund, 2010; Goldin, Katz, & Kuziemko, 2006; Duckworth & Seligman, 2006; Hyde & Mertz, 2009; Hyde, Lindberg, Linn, Ellis, & Williams, 2008; Fortin, Oreopoulos, & Phipps, 2015)[1] women still earn on average around 16% less per hour than men in the EU (The European Commission, 2014) and around 18% less in the US (The US Bureau of Labour Statistics, 2014).[2] The reasons for these wage and education gaps are complex and manifold.[3] Access to higher education is one important prerequisite for later employment possibilities and wages and is in many countries determined *inter alia* by university entrance exams.[4] These exams often use multiple-choice testing formats—especially in the US—because it is considered to be efficient, it allows for large-scale testing and for a broad coverage of content (Frederiksen, 1984).[5] Nevertheless, multiple-choice test-

* Corresponding author.
E-mail addresses: riener@dice.uni-duesseldorf.de (G. Riener), wagner@dice.uni-duesseldorf.de, wagner@dice.hhu.de (V. Wagner).

[1] Goldin et al. (2006) show that females gained about 0.17 of a standard deviation from 1972 to 1992 in standardized math tests in the US and Hyde et al. (2008) show that gender differences in mathematics skills are close to zero for grades 2–11. Hyde et al. (2008) analyze scores on the National Assessment of Educational Progress (NAEP) of about 7 million eighth graders of 10 states in the US. In contrast, using data from the Early Childhood Longitudinal Study Kindergarten Cohort, Fryer and Levitt (2010) find that there are no mean gender differences upon entry to school in math standardized test scores, but that girls lose more than two-tenths of a standard deviation relative to boys over the first six years of school.

[2] Furthermore, the wage gap between skilled and unskilled population groups has been rising since—at least—the 1970s (see Marquis, Trehan, and Tantivong, 2014, and the literature mentioned therein).

[3] According to the European Commission possible explanations could be inter alia discrimination in the workplace, different jobs in different sectors (STEM fields), the undervaluing of women's work and skill and women's under-representation in senior and leadership positions.

[4] In the US, the weekly earnings of women with only a high school diploma represented 83% of the earnings of women with an associate's degree and 55% of the earnings of women with a bachelor's degree or higher (The US Bureau of Labour Statistics, 2014).

[5] Also, in Germany multiple-choice questions constitute an important testing format in centralized comparison tests (VERA, PISA, TIMSS) and in university exams. Furthermore, the testing of cognitive knowledge also predicts and correlates well with overall competence and performance (McCoubrie, 2004).

http://dx.doi.org/10.1016/j.econedurev.2017.06.005
0272-7757/

ing formats are not without problems if they favor the answering strategies of certain groups in the population.

The analysis of gender differences in standardized multiple-choice tests with respect to *performance* (Jurajda & Münich, 2011; Ors, Palomino, & Peyrache, 2013) or *skipping test items* (Akyol, Key, & Krishna, 2016; Ben-Shakhar & Sinai, 1991; Pekkarinen, 2015) has therefore received some attention. Recent experiments have identified skipping—women tend to skip more items than men—as one reason for men outperforming women (Baldiga, 2014; Pekkarinen, 2015), but as promotion within the educational system should depend on actual knowledge and not on how knowledge is assessed, this poses a challenge for general multiple-choice tests. The negative effect of skipping test items on performance has also been recognized by the College Board which in March 2016 redesigned the scoring rules of the SAT. The old scoring rule—students get $\frac{1}{4}$ points deducted for an incorrect answer—has been changed to a "rights-only" scoring method. Under the new scoring rule, students receive one point for each correct answer and each incorrect answer receives zero points "*to encourage students to give the best answer they have for every question without fear of being penalized for making their best effort*" (The College Board).[6] Structural biases in multiple-choice testing would therefore challenge the use of this testing format and understanding the underlying causes is important, particularly if differences are driven due to a higher willingness to skip questions and not due to differences in ability.[7]

Although any testing format is advantageous for some and less advantageous for others (e.g., oral exams could favor extroverts; open-ended questions could favor females...), taking all characteristics of the general population into account in the design of tests for promotion is important in order to at least not increase educational inequalities and to develop an institutional setting for equal opportunities in education. However, standardized test scores are used for placements and admissions at nearly every level of schooling (Baldiga, 2014) and it is therefore important to know for which subgroups of the population gender differences emerge and how these differences could be mitigated.

So far, researchers have mainly focused on gender differences in university entrance examinations and have hence relied on a highly selective student subject sample which does not allow us to analyze differences in skipping for a more heterogeneous population (Baldiga, 2014, on Harvard Business School students, Espinosa and Gardeazabal, 2013, on second year undergraduate students, Akyol et al., 2016, on the Turkish university entrance exam and Pekkarinen, 2015, on Finnish students taking university entrance exams).[8] There is little evidence on school-age children although this is important for progressing within the school system. Moreover, these types of tests may determine the transition between school types and may therefore determine pupils' future educational success.

In this study, we analyze gender differences in skipping test items in a mathematical test for fifth and sixth graders in secondary schools in Germany. Our sample—although not representative—tries to closely mirror the general population and is one of the first among *school-age* children who either attend a vocational school (preparing for blue color occupations) or a high school (preparing for higher education). The German school system is particularly well-suited for studying answering strategies for heterogeneous groups of the population as the transition decision in Germany is strongly related to parents' socio-economic background (Dustmann, 2004). Pupils are segmented into the different school types at the age of 10. There is little mobility across school types after initial tracking. Consequently, school types differ to a large extent in pupils' social background and pupils' intellectual ability. For instance, pupils at a vocational school report having significantly fewer books at home than pupils at a high school (26–100 vs. 101–200, $p<0.01$), but we cannot rule out that school types also differ in other dimensions, such as teacher quality. Nevertheless, we reduced the influence of the school type to a minimum by conducting our experiment in the first two years of secondary school. Hence, our findings between school types can be seen as suggestive evidence of differences in socio-economic background.

We applied questions from a German-wide mathematics competition test (*Känguru-Wettbewerb*)[9] of which we obtained reliable data on item difficulty.[10] We varied the test item difficulty within subjects and increased the attractiveness of answering difficult questions. The expected value from answering completely at random is negative for easy questions, zero for medium questions, but positive for difficult questions. This gives us a strong test of the persistence of the gender gap for difficult items when answering at random is made more attractive. Moreover, we use external incentives to increase the stakes of the test; pupils received non-monetary rewards for improving over their own previous mathematics results.[11] We expect extrinsic incentives to affect girls and boys differently if these incentives increase the stakes of the test. Azmat, Calsamiglia, and Iriberri (2016)—in a recent study—have shown that the gender gap in *performance* varies with the stakes of a test. As we could not vary how much the test counts toward the final grade, as in Azmat et al. (2016), we opted for non-monetary extrinsic incentives. This allows us to test whether the gender gap in *skipping* also exists if the stakes are low. Previous studies have focused only on high stakes university entrance exams. We can determine whether the gender gap changes if the item difficulty is less salient by setting the focus on winning a prize.

In our—non-incentivized—baseline treatment, we find that girls skip more questions than boys but only if the test items are difficult. Interestingly, gender differences in skipping are no longer detectable and are small if pupils can win a reward. Differentiating by school type, we find that this gender gap can be attributed to pupils in high schools. Our findings in high schools are in line with the findings in the literature but the non-existence of a gender gap in vocational schools challenges a conclusion that this gap observed in high schools is driven by innate differences by gender. This also constitutes a challenge to the external validity and shows the importance of taking into consideration all social levels of a population.

To the best of our knowledge, this is the first study using a secondary school sample and that compares the gender gap in skipping between pupils of different school types. Moreover, there are only a few studies using data from framed field experiments (Baldiga, 2014; Espinosa & Gardeazabal, 2013) and no randomized field experiment has been conducted so far in secondary education.

Recent literature documents that girls outperform boys in terms of GPA but that boys still perform better on standardized tests

[6] See https://collegereadiness.collegeboard.org/pdf/test-specifications-redesigned-sat-1.pdf.

[7] Other reasons for the answering gap in (mathematical) multiple-choice tests could be women's retention for competitive settings (Niederle & Vesterlund, 2007; 2010) or the stereotyping that women perform worse in mathematics than men.

[8] The study by Ben-Shakhar and Sinai (1991) is one exemption using inter alia data on ninth graders in Israel.

[9] http://www.mathe-kaenguru.de/wettbewerb/.

[10] We could not vary the order of item difficulty—easy to difficult—as this was also predetermined by the *Känguru-Wettbewerb* and was a prerequisite for schools participating in the study.

[11] Phelps and Price (2016) use extrinsic financial incentives among elementary pupils and allow pupils in the treatment group to publicly signal when they have finished all test questions. They show that pupils derive social rewards of completing the task faster than their classmates but that test performance is lower.