I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11202

Ordering History Through the Timeline

Eugenio Garibaldi
Pietro Garibaldi

DECEMBER 2017

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11202

Ordering History Through the Timeline

Eugenio Garibaldi
Bocconi University and Garycom s.r.l.s.

Pietro Garibaldi
Collegio Carlo Alberto, University of Torino and IZA

DECEMBER 2017

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Ordering History Through the Timeline*

History is a key subject in most educational system in Western countries, and there is ongoing concern about the the degree of historical knowledge and historical sensibility that students obtain after their high school graduation. This paper proposes a simple *linetime* test for quantitatively measuring a human sense of history. The paper reports the results of the test administered to approximately 250 Italian university students. There are two empirical results. First, students have remarkable difficulties in ordering basic events over the time line, with the largest mistakes observed around the events that took place in the Middle Age. Second, the paper uncovers a statistical regularity in the test performance across gender, with female subjects featuring a statistical significant and quantitatively sizable downward score. The gender difference is surprising, since existing literature on differences in cognitive abilities across gender suggests that female subjects outperform male subjects in memory related tests. The paper shows also that the gender difference survives to a variety of sub periods, and falls by only 20 percent when we distinguish between violent and non violent events.

JEL Classification: I12
Keywords: sense of time, gender differences

Corresponding author:
Pietro Garibaldi
Collegio Carlo Alberto
Via Real Collegio 30
10024 Moncalieri (Torino)
Italy
E-mail: pietro.garibaldi@carloalberto.org

* We thank Alberto Anfossi and Compagnia di San Paolo for financial support and Elena Ovazza for inspiring us to pursue this research.

1 Introduction

The sense of time in history is an individual ability to place historical events through the timeline. In contemporary societies, educational systems devote an incredible amount of effort and resources to study the unfolding of key events over the last 2000 years. In some countries, and Italy in particular, history is studied for 13 years by every single student, independently from her/his high school specialization. In our perspective, the sense of time belongs to history in the same way as addition and multiplication belong to algebra and mathematics. Arguably, perception of the historical distance of key events is important across most social sciences. Not surprisingly, individual time perception is an entire field of study within psychology, cognitive linguistics and neuroscience, and refers to the subjective experience of time and its measurement (Grondin, 2010).

Unfortunately, we know very little about individual and student ability to place events along the timeline. Most international educational tests focus on math, science, reading, verbal literacy, and even financial literacy (see OECD 2016), but do not spend time and resources to measure the historical sense of time. In addition, standardized tests on history, such as the SAT in the US, do not focus on measuring the sense of time, but are based on broad multiple choice questions. There is media concern about the actual knowledge of students (HuffPost, 2011), and scholars are actively proposing alternative ways of teaching history (Davies, 2010, VanSledright, 2011)

Measuring objectively the sense of time is not trivial, since it is akin to measuring a basic skill in a humanity oriented field. The present paper has two objectives. First, it provides a simple test for measuring quantitatively a human sense of history without relying on any specific mnemonic date. This *linetime* test contributes to quantitatively measuring basic knowledge in humanities, and can be applied to many disciplines. Second, it presents the results obtained by administering the tests in experimental setting to some 250 Italian students. There are two empirical results. First, students have remarkable difficulties in ordering basic events over the time line. Second, the paper uncovers a statistical regularity in the test performance across gender, with female subjects featuring a statistical significant and quantitatively sizable downward score.

The *linetime test* that we provide turns out to be particularly simple. It comprises of three key steps. First, we identify N key historical events over the timeline. While the relative importance of key historical events obviously varies from country to country, our current research focuses on the case of Italy and we select 32 key events that unambiguously do not overlap. Note that a key feature of the *linetime tests* is that events are never defined or classified by a specific date. Second, we randomly mix the events and ask subjects to ordering these events through their perceived timeline. Third, and finally, we propose a quantitative metric distance between the subjective ordering and the actual objective ordering, and we thus map the individual performance in the *linetime test* into a single natural number.

In the second part of the paper we describe the results of the *linetime* test administered in an experimental setting to 250 university Italian students in economics. The results show a sizable variance in the performance of the *linetime test.* The average student is generally capable to place historical events that occur at the tails of the time line. In other words, events such as the Egyptian pyramids and the fall of the Berlin Wall are correctly positioned at the far end of the perceived timeline. Yet, events that occurred in the Middle Age, such as the fall of the Byzantine empire or the 100 Years War between France and England appear largely obscure to the average student. The average event is miss-placed by two positions. We thus legitimately wonder whether the teaching of history is satisfactory vis-à-vis the Italian Ministry recommendation that students acquire "a sense of history".

Existing literature on differences in cognitive abilities across genders suggests that female subjects outperform male subjects in memory related tests (Halpern, 2012). Perhaps surprisingly, the experimental results of our *linetime test* feature a sizable difference across genders that survive various robustness checks. Our research shows that female subjects perform on average below male subjects by as much as 50 percent of the standard deviation of the entire sample and of the male subjects. We perform various econometric exercises to understand the gender difference. First, we run a battery of multivariate regressions, since we also have information on student performance in the national high school exam (so called "maturità" test), on their high school grade, and on the their family background. Our regressions show that the gender differential in the *linetime test* is as large as twice the standard deviation of the results in the national "maturity test". Second, we divide the line time test in historical sub periods (Ancient time, Middle Age, Contemporary History, and World War II and Cold War), and find that the gender difference is present in each of the four sub periods. Finally, we also distinguish between violent and non violent events, and find that that the gender differential is still highly significant, although it reduces by some 20 percent.

The paper proceeds as follows. We first briefly review the literature on the teaching of history and on the existing standardized tests, as well as the vast literature on the gender differences in cognitive abilities. Section 2 also defines the aim and scope of our analysis with respect to these literature. Section 3 defines the linetime test in its general form, and provides the formal metric that we propose. In addition, section 3.2 studies the statistical features of the linetime test obtained by montecarlo simulations based on fictitious agents that randomly play the test without any historical knowledge. Section 4 outlines the experiment carried out in the case of Italy while Section 5 reports the outcome of the experiment and the distribution of mistakes across the timeline. Section 6 reports the results of the multivariate analysis, and the robustness of the gender differences. Section 7 concludes.

2 Literature Review and scope of the analysis

Our paper is related to at least two strands of literature within the social sciences. Firstly, the undergoing research on assessing and proposing alternative methods to test historical knowledge. Secondly, the literature on gender differences in cognitive abilities. While we briefly touch upon these two strands of literature, we also clarify the scope of our analysis.

To the best of our knowledge, the *linetime test* is a new quantitative method for measuring historical competences. Standardized historical tests are part of the SAT exam in the US educational system, but they are typically based on broad multiple choice questions. Such tests are certainly a genuine method for measuring historical knowledge, yet they are subject to intense scrutiny. There is a strong debate among educational scholars on how history should be taught. Huijgen et al. (2014) claim that the formal education system is based too much on mnemonic items and does not develop a historical perspective taking (HPT). Huijgen et al. work is based on a test proposed by Hartmann and Hasselhorn (2008). The latter is based on a case study approach to history. For example, to understand why a German citizen could have voted for the nazi party in the 30's, students are presented with a real life scenario of a particular individual living in Germany during the Weimer republic. Huijgen et al. argue that a case study approach will help students developing a deeper understanding of the causes and motivation behind key historical episodes. Independently of the taste for a case study approach, the *linetime test* that we propose is substantially different. VanSledright (2004, 2011) is also fairly critical about the way history is taught in the US, and argues that too much effort on memorizing leads to a poor knowledge among students. He claims to develop an assessment method that helps developing a critical mind, but ultimately his tests remain a combination of multiple questions and essay-type questions.

We are not claiming that the proposed *linetime test* will help students into a new critical thinking about history. Yet, we believe that a simple quantitative test on the ability to place events through the timeline should be within the toolbox of the educational system. Correctly placing events into the timeline is akin to performing numerical operations in basic math. Once students will be comfortable with the basic timeline, they will be in a better position to develop a critical historical sense.

The second strand of literature that we aim at contributing to is the huge work on gender differences in cognitive abilities. Such literature attracts attention of most social sciences, including economics, psychology, evolutionary psychology as well as neuroscience. A key reference in this respect is Halpern (2004, 2012). She claims that social scholars should be interested not in *who* is better between males and females in certain cognitive tasks, but rather on finding *"where and when meaningful differences are found"*. Psychologists also claim that uncovering differences in cognitive abilities will help people achieving their maximum potential.

With respect to the differences in cognitive abilities observed by psychological tests, the *linetime test* is likely to interact with at least two empirical regularities. First, female subjects perform persistently better in memory related tests. Second, male subjects perform better in tests that deal with visuospatial abilities. We look at the first regularity. Ordering events through the line time certainly requires some form of memory.

Yet, memory is a multi dimensional concepts. Female subject performs better in episodic memory (memories for which someone can remember where and when they learned the information being processed), as well in autobiographical memory (memories that pertains to events in one's life). It seems that female subjects have also better performance in tests related to short term memory (i.e. for events that took place in the previous 1 or 2 minutes), as well as memories for odors. The latter has arguably little to do with the abilities tested in the *linetime test*. Nevertheless, female subjects have also better memories for spatial locations, or the ability to remember if an object was seen in a room or where exactly an object was seen in a photo. In this respect, it is somewhat surprising to uncover that female subjects perform worse than male subjects in our test.

The second set of abilities that the *linetime test* is likely to require relates to visuospatial abilities, an individual skill in representing, transforming, generating symbolic and non linguistic information. These type of skills are also multivariate, and refer to spatial perceptions (ability to locate the horizontal or vertical in a flat display), mental rotation (ability to immagine rotated objects), spatial visualization (tap spatial visualization), and spatiotemporal ability (judge the dynamic of moving image). Psychologists argue that these type of skills are used in engineering, architecture, chemistry, and in many hard sciences. In this type of skills male subjects outperform female subjects. It is far from straightforward understanding why ordering historical events - in a way similar to what is required in the *linetime test* - must be correlated with these type of skills. Yet, since these cognitive differences are large in the literature and the differential is large, the *linetime test* may partly operating along these lines.

Beyond uncovering the regularities, the psychological literature speculates on whether the gender differences are driven by nature of by nurture. Half of Halpern (2012) excellent work is devoted to these possibile explanations. Arguing that gender differences are biological is considered dangerous, since it may end up justifying gender inequality in modern societies. In addition, it is well established that the environment influences gene expression (Benbow and Stanley, 1980). Evolutionary psychology- a field in the middle of psychology and biology- proposes the hunter-gatherer hypothesis for understanding cognitive differences. While evolutionary theory is certainly appealing, it can also be misleading, since there is no experimental evidence to identify the right channel. According to the hunter-gatherer hypothesis, the better performance of female subjects in memory related tests reflects the fact that in primitive societies females needed good memory for locating plants in their role as gatherers. At the same time, male needed better visuospatial abilities to specialization in hunting related activities.

Finally, there is also a more economic oriented strand of literature that shows how females shy away from competitive settings, and such behavior may explain some differences in test scores. Niederle and Vesterlund (2010) argue that competition plays a key role for the gender differences in math test scores. In our test students are offered small incentives to participate and perform, and the gender differential may partly due to different level of attitudes towards competition across gender. Kautz et al. (2017) strongly argue that measures of cognitive skill are sensitive to incentives. They claim that test scores for young children can be improved by one standard deviation by offering candy for correct answers. It may well be that male subjects respond more to these basic incentives. In addition, the performance in our test has also to do with the selection of people into a subject like history. McDonald et al. (2012) propose the "male warrior hypothesis", a psychological theory whereby men's psychology is designed in ways that facilitate success in intergroup conflicts. In this sense, the aggressive attitude would lead male to be more interested in a subject like history where the narrative of intergroup conflicts is the bread and butter of the discipline.

Our paper does not aim at entering the details of these psychological debates, and we do not aim at offering definite explanations. Our ultimate aim is to provide an original test on historical knowledge, to show the experimental results in the case of Italy, and to document the gender differences that we uncover.

3 The Linetime test

The *linetime test* is carried out in three key steps. First, we identify N key historical events over the timeline that unambiguously do not overlap. In selecting and defining the key facts of the *linetime tests*, dates are never mentioned in the test. Some care must be used to avoid events that somewhat overlap. For example, the events "World War II (WWII)" and "Pearl Harbor Attack" can not be both used for the test since the Pearl Harbor attack took place during World War II, and there is no unambiguous ordering of the two historical facts. Conversely, the events "German invasion of Poland" and "Pearl Harbor Attack" are legitimate candidates, since the Nazi invasion of September 1939 unambiguously took place before the Pearl Harbor Attack of December 1941.

Once the N events have been selected, the subject is presented with the events randomly mixed. This is the second step. The task of the test is simply to order the events according to her/his perceived timeline. In the experiment carried out in the next section, subjects had up to 30 minutes of time for ordering 32 events through the timeline. The details are explained in the next section.

The last and third step of the test requires a scoring method for assessing the distance between the perceived timeline of the subject, and the objective ordering of the events as they happened through history. More formally, one needs a metric distance between the subjective ordering and the actual objective ordering, so as to summarize the performance in the *linetime test* by a a single natural number. While this last step is a bit more formal, the intuition of the scoring method we propose is very simple. We consider a penalty score for each event that is equal to the absolute distance between the objective ordering and the subjective ordering. In other words, if the event in the Nth position is placed in the first position by the subject, the penalty for that event is equal to N. The final scoring of the test is the sum - in absolute value - of the total penalties. In the next subsection we describe the formal definition of the scoring method.

3.1 The scoring method

Let H be a collection of N historical events ordered through their actual happening over the time line. Formally, H is defined as

$$H = \{h_1, h_2, ..., ..., h_N\}$$

where h_m is the position of the historical event m, with $m \in H$. Note that $h_m < k_{m+j}$ implies that event h_m took place before event $m + j \quad \forall j > m$. In other words, the set H records the objetive ordering of the events.

Let J^i be the guess of the ordering of the N events in H played by individual i. We can thus say that J^i is the subjective timeline of individual i of the N events over $his - her$ timeline so that

$$J^i = \{j_1, j_2, ..., ..., j_N\}$$

where $j_k < j_{k+j}$ implies that individual i believes that event j took place before event $k + j$, $\forall j > k$. A general scoring method is a distance between the historical ordering H and the guess J^i. A general penalty can be defined as

$$LT\_SCORE(J^i) = -\lambda(J^i),$$

where λ is the penalty associated to guess J^i of individual i. In the *linetime* test of this paper the penalty λ reads

$$\lambda(J^i) = \sum_{i=1}^{N} |j_i - h_i|$$

so that the $\lambda(J^i)$ is the sum of the absolute values of the difference between the position of event i in H- that we label h_i- and the position of event i in J^i, that we label j_i. The event line time score is thus

$$LT\_SCORE(J^i) = -\sum_{i=1}^{N} |j_i - h_i|$$

The function $\lambda(J^i)$ is highly non linear, and gives much more weight to larger than to smaller mistakes. Switching the position of two consecutive events implies a penalty equal to 2. Conversely, placing the N^{th} events in the first position yields a penalty of $2N$, since beyond the penalty of N for the mistake itself, there is bandwagon effect equal to N, as all events end up wrongly placed by one position. Note that in the experiment carried out in Italy we rescale the total score by 100, so that we work with the function $\overline{LT\_SCORE(J^i)} = 100 - LT\_SCORE(J^i)$. Somehow we feel that that subjects that obtained a perfect matching are "happier" with a score of 100 rather than a score of 0.

Table 1: **The *Linetime* test played in Italy: Historical events**

	Italian name	**English translation**	V/NV	Sub_P
1	Piramide di Cheope	Pyramid of Cheope	NV	A
2	Micene e Maschera di Agamennone	Mycenae and Agamemnon Mask	NV	A
3	Pericle e Partenone	Pericles and Parthenon	NV	A
4	Alessandro Magno	Alexander the Great	NV	A
5	Guerre Puniche	Punic wars	V	A
6	Giulio Cesare e Augusto	Julius Caesar and Augustus	NV	A
7	Editto di Costantino e Cristianizzazione dei Romani	Edict of Constantine and Christianity of the Romans	NV	A
8	Caduta Impero Romano di Occidente. Sacco di Roma	West Roman Empire's Fall. Sack of Rome	V	A
9	Incoronazione Carlo Magno	Charlemagne- coronation	NV	MA
10	Marco Polo e Repubblica di Venezia	Marco Polo and the Republic of Venice		MA
11	Guerra dei Cent'anni (Francia e Inghilterra)	Hundred Years War (France and England)	V	MA
12	Caduta Impero Bizantino. Presa di Costantinopoli	Byzantine Empire Fall. Grip of Constantinople	V	MA
13	Rinascimento. Morte di Lorenzo il Magnifico	Renaissance. Death of Lorenzo the Magnificent	NV	MA
14	Lutero e Nascita Protestantesimo	Luther and Birth Protestantism		MA
15	Guerra dei 30 anni	30 Years War	V	MA
16	Regno di Re Sole	Kingdom of the Sun King	NV	MA
17	Rivoluzione Francese	French Revolution	V	CT
18	Napoleone- Incoronazione	Napoleon- Coronation		CT
19	Congresso di Vienna	Vienna Congress	NV	CT
20	Risorgimento. Unita' d'Italia	Risorgimento. Italy's Unification.	NV	CT
21	Assassinio di Sarajevo	Assassination of Sarajevo	V	CT
22	Rivoluzione Russa	Russian revolution	V	CT
23	Marcia su Roma	March to Rome	V	CT
24	Guerra Civile Spagnola	Spanish Civil War	V	CT
25	Invasione della Polonia della Germania	Germany invades Poland	V	CW
26	Battaglia di Londra (Luftwaffe-RAF)	Battle of London (Luftwaffe-RAF)	V	CW
27	Sbarco in Normandia	landing in Normandy	V	CW
28	Resistenza e 25 Aprile	Resistance and April 25th	V	CW
29	Bomba Atomica Hiroshima	Atomic bomb Hiroshima	V	CW
30	Guerra di Corea	Korean war	V	CW
31	Guerra di Vietnam	Vietnam war	V	CW
32	Caduta Muro Berlino	Fall of the Berlin wall	NV	CW

The third column V/NV states if the event is Violent (V) or Non Violent (NV) according to our classification.
Sub_P refers to sub-periods:
A refers to Ancient, MA refers to MiddleAge, CT refers to Contemporary, and CW refers to WWII and Cold War

3.2 The statistical properties of the score $LT\_SCORE(J)$

In the test that we carry out below the number of events N is 32. This implies that the possible realizations of the $LT\_SCORE(J^i)$ are 32!, corresponding to the permutations of 32 numbers. In other words, the possible random ordering of the 32 events is in an order of magnitude of 10^{35}, a remarkable large number. This suggests that a subject that plays the *linetime* test by random guessing has a probability of perfect score not far from zero. Further, if we use a value of 100 for a performance with zero penalty, as we do with the score $\overline{LT\_SCORE(J^i)}$, it is very unlikely that an agent that randomly plays the *linetime* test without any prior knowledge of history gets a positive outcome.

To investigate the issue further, we simulate agents that randomly play the *linetime test*. We immagine having 10^6 agents that are totally clueless about history, but understand the rules of the game and thus play randomly. We then record their score. The results are plotted in Figure 1. The average score obtained by the random players is approximately -241 while the standard deviation is 38. The empirical distribution is clearly symmetric. The minimum score obtained is around -380 while the maximum is -80. We also replicate the experiment 300 times to check whether we obtain a a positive outcome. It never happened. The best outcome we ever got by random players was -20.

4 The Experiment in the case of Italy

We asked economics students at the University of Torino to take part to a "test on humanity and social issues". To participate to the test students had to be enrolled in economics and not in business. Potential subjects were offered 5 euros in food coupons as a participation incentive, and had to subscribe by email. Students were also informed that an additional 5 euros in coupon were offered for good performance in the test. Students were thus completely clueless about the content of the test. Approximately 250 students replied on line to the offer of participation.

The events selected for the experiment are listed in Table 1. These include wars (from the Punic wars of the Roman empire to the Korean war of Worlds War II), famous historical people (from Alexander the Great to Marco Polo, and Napoleon) and other broad categories (Egyptians Pyramids, the Vienna Congress, the fall of Berlin wall...). All of the events are unambiguously studied in details during the compulsory Italian education system. Generally speaking, the choice of the event is country specific. In the case of Italy we include some important events with similar names, to check whether students did confuse the two. As an example, we included both Alexander the Great and Charlemagne. In real life, the two "Great" persons actually lived more than a millenium apart. Further, we include three important events of Italian history that may sounds similar, albeit very distant from each other. The three events are the 16th century Renaissance, the Risorgimento that lead to a unified Italy in 1861, and the so called Resistance, the strong opposition to the late Mussolini fascist regime from 1943 to 1945. Note that these events have very little in common, beyond the fact that all start with the same alphabetical letter "R" . Even though there is a bias on Italian history, we selected key events from the entire world history. We are not claiming that they are the most important events, but they are unambiguously taught in the Italian curriculum. To make the test somewhat challenging, we included also the '"30 years war", arguably the most dramatic historical events in in the history of the seventeenth century. As we describe in Table 1, we classify *ex-post* (without obviously telling it to the students), 4 sub periods of 8 events each. This implies that the ancient period goes from the Egyptian Pyramids to the fall of the West Roman empire. The MiddleAge moves into the 16th century and up to the reign of Louis XIV (known as Re Sole in the Italian tradition). The contemporary period runs until events the took place before WWII, and in our selection they stop with the Spanish Civil War. The last block of events is labelled WWII and the cold war, and goes from WWII until the fall of the Berlin wall. Note that in the 20th century there are many events for a relatively short historical period. This turned out to be a way to test whether students could order what happened within a relative short and dramatic period in Italian history, such as WWII and the end of the fascism.

On the test day 246 students showed up. The test took place in 3 phases. In the first phase students were asked to provide detail information about their high school grade in history, their grade at the national high school exam (so called "maturità" in the Italian system), the type of high school they attended as well as information on the educational background of their family. Finally, students assessed the number of books they have within the households. In the second phase of the test students were introduced to the rules of the *linetime* test without having access to the list of events. One student left the room at this stage. Finally, the events were distributed to the students with different metric to avoid potential cheating. At that point the test began and students were given 30 minutes to order the events. Students had to hand in the test to qualify for the 5 euros participation reward. The bonus was given to 50 best students once the results were compiled, and students were notified of their good score.

Table 2 reports main summary statistics of our key variables: score in the test, history grade and high school grade. Few observations are immediate. First, the average score in the sample is 36.9 and the standard deviation is larger than the mean. One subject obtained a score as low as [-168] while 9 students (approximately 2.5 percent of the sample) obtained the full mark of 100. Figure 2 reports the entire distribution of the score. In the left tail, 5 students have grades below -100, and 41 students obtained a score below 0, corresponding to 16.6 percent of the sample. From Figure 2 few properties are immediate. First, the distribution is not symmetric since it is truncated in its right tail at the maximum grade of 100. Second, the distribution is very different from the theoretical distribution simulated by random players in Figure 1. Not surprisingly, students do not play randomly.

Summary statistics on the rest of the variables are reported in Table 2. The average history grade ranges

Table 2: Summary statistics

	All - mean	Male - mean	Female - mean	All-Range
Score in timeline	36.9	50.9	21.7	[-168,100]
	(50.7)	(52.1)	(46.7)	
History grade	8.1	8.1	8.1	[6,10]
	(0.9)	(0.9)	(0.9)	
Final high school grade	81.4	80.1	83.3	[60,105]
	(11.4)	(11.3)	(11.4)	
Family Education	26.3	27.2	25.1	[10,44]
	(6.7)	(6.7)	(6.6)	
Number of Books	2.9	3.1	2.7	[0,6]
	(1.8)	(1.9)	(1.8)	
University Year	1. 7	1.8	1.6	[1,3]
	(0.76)	(0.8)	(0.7)	
Classical High School	0.14	0.12	0.16	[0,1]
	(0.35)	(0.3)	(0.37)	
Scientific High School	0.42	0.49	0.33	[0,1]
	(0.5)	(0.5)	(0.47)	
Foreign Language	0.08	0.05	0.12	[0,1]
	(0.3)	(0.2)	0.4	
Observations	244	136	107	

Standard Deviation in Parenthesis

from 6 to 10, reflecting the Italian grading system in high schools. The Final high school grade- conversely- ranges from 60 to 105. Note that we assign 105 to students that obtained the full grade of 100 *cum laude*. Family background is measured in years of education, and students were asked both the education of the mother and the father. The range goes from 10 to 44. The number of books in the household is a numerical variable that is constructed on the basis of the questions asked to subjects. The university year corresponds to the year of enrollment of the student. Finally, the Table reports information on the proportion of students that speak a foreign language at home. Column (2) and (3) of Table 2 reports the same information across gender. These differences are analyzed in details in Section 6.

Table 3: Mistakes by sub-periods

Subperiod	Average Mistake	Least known	Most Known
Ancient history (1-8)	1.47	Alexander the Great	Pyramid of Cheope
Middle Age (9-16)	2.56	Hundred Years War	Reneissance
Contemporary history (17-24)	1.82	Vienna Congress	Assassination of Sarajevo
World War II and Cold (25-32)	1.49	Battle of London	Atomic Bomb of Hiroshima

Average mistake is the mean error associated to a particular event. It is obtained by averaging all the penalties associated to one event across all the sample.

5 How much students know?

In the Italian system history is taught for at least two times during the compulsory education. The Italian education system is centralized, and we can check the formal objectives given by the Ministry of Education (*D.M. 9 febbraio 1979*) to the each school and teacher. We focus on "middle schools" (i.e. schools in grade 6th to 9th attended by pupils aged 11 to 14). The Ministry argues that the primary goal should be to acquire a "sense of history". Moreover, students should understand past societies and develop a temporal dimension in their way of thinking. This include knowing and remembering the most important historical events. Nevertheless- the Ministry recommends- the teaching of history should not be based on a "encyclopedic knowledge". The syllabus and the government instructions are further divided across the three years of the "middle schools". In grade 6th (first year of the middle school), history should range from the middle age to 1492 (with some hints to the ancient period of greeks and romans). In grade 7th the syllabus should range from 1492 to the end of nineteenth century, and in grade 9th teachers should cover the history of the 20th century up to the the present days. Particular emphasis should be given to Italian history. After the middle school, students are expected to enter into different high schools for at least two years (the compulsory education in Italy is up to 16 years old). Broadly speaking, the *Liceo's* are the most academic oriented high schools, while the technical schools should be more oriented toward professional specialization. Independently of the high school selected, all students take 5 more years of history, and the curriculum once more goes from the ancient period to the present times. After 5 years of high school, students take the national "maturità" and enroll in the university.

As mentioned above and in Table 2, the average score in the exam is 36.9. This implies that the average $\lambda(J^i)$ penalty of our undergraduate student is 63.1, since we normalized the full score to 100. The result is somewhat remarkable. First year undergraduates have just completed their high school and should have fresh memories of events studied for at least two times during their compulsory education. In light of the goals set forward by the Ministry, it is difficult to argue that competences were acquired satisfactorily. The *linetime* tests shows that these students have great difficulties in ordering basic historical events. An average penalty of 63 implies that students on average miss two positions in their perceived timeline respect to the objective ordering. Just to give few examples, one third of the students does not know that the most recent historical event is the fall of the Berlin wall. Moreover, only a 7% of the sample does less than two small mistakes. One may legitimately wonder what the average students was taught and was expected to know in order to complete his/her high school.

Figure 3 plots the distribution of error for 32 historical events. While the average mistake is approximately 2, the distribution has an interesting bell shape. This implies that students make smaller mistakes at the far end of the distributions, and are generally capable of knowing that the Egyptian Pyramid are an ancient event while the fall of the Berlin wall took place in the present days. The largest mistake refers to the Hundred years War (event 11), a key MiddleAge event in Western Europe. Students are also fairly clueless about the actual position of the Vienna Congress of 1814, when Europe decided on a new order and on the restoration of sovreign kingdom in the aftermath of the Napoleon era. Table 3 reports the average mistakes by sub-period, and confirms the eyeball finding of Figure 3. The largest penalty was obtained by events in the MiddleAge, with an average mistake equal to 2.56.

6 Gender differences

The interesting- and pheraps surprising- finding of Table 2 refers to the the average score obtained by female and male students. Whereas male students obtained an average score of 50.9, the average score of female students is 25.6. To give a sense of the average difference relative to the standard deviation, Table 4 reports simple test of mean differences. The so called *d-test* is popular in the psychology literature (Halpern, 2012), and a value of the d-statistics larger than 0.5 is typically considered very large. In Table 2 the d-statistics is 0.53. Since the underlying distribution is not symmetric nor normal, the d-statistics in Table 4 have been bootstrapped. Figure 4 plots the empirical distribution of grades by gender, and shows that the two distributions do not overlap, and very few female subjects obtain grades in the right tail of the grade spectrum. This differential in very high grade is coherent with the general findings of Hedges and Novell (1995).

The large gender difference observed in Tables 2 and 4 deserves to be further analyzed through multivariate regressions. As reported in Table 2 we have various potential controls that can help us to understand the gender differences. Table 2 reports also summary statistics on the main explanatory variables by gender. The history grade in the last year of high school is almost identical across gender, while the high school grade (the "maturità" grade) is higher for female than for men, respectively 83.3 versus 80.1. According to national averages compiled by Almalaurea (Corriere della Sera, 2016) the average national grade is 78.4 for female and 75.2 for male. This suggests that the sample of our students is above the national average but keeps the gender differential observed in the national data. With respect to selection into high schools, female subjects are over represented among the classical high school and are less represented in the scientific curriculum. There are also more female that speak a foreign language at home. Finally, the share of female in the sample is 43 percent, a percentage very similar to the share of female enrolled in economics at the University of Torino.

We run the following simple regression model for the score of individual i,

$$Score_i = \alpha + \beta Gender_i + \gamma HighSchool_i + \delta History_i + \sum_{j=1}^{n} \rho_j X_j + \epsilon_i \tag{1}$$

The coefficient of the gender difference is β. The vector X_j includes controls other than the high school "maturità" and the history grade. The results are reported in Table 5 for 3 different specifications of equation (1). In the first column we include only the female gender dummy, and the coefficient reflects the simple average score differential in the sample. The gender difference is not reduced when we control for high school grade and history grade (Column 2), even though they are both correlated with the score statistic. Column (3) shows that the β coefficient on gender differences is basically unchanged even when we control for all the variables in the dataset. To give a further sense of the magnitude of the β coefficient, column (3) implies that the gender differential is as large as twice the standard deviation of the results in the national "maturity test". In the next two subsections, we further look into this gender differences.

6.1 In which historical periods is the gender difference arising?

One may wonder whether the sizable estimate of the β coefficient arise in any particular sub group of periods. We want to check whether male subjects have a natural attitude to order events in a particular interval of the timeline. We thus now assume that each subject played different sub games of $M < N$ events, where M is the number of events of the sub-game , while N is the number of events of the original game. Further, we assume that each sub-game is played independently of the other sub-games, and we record its relevant grade.

We thus divide our 32 historical events into the 4 historical periods mentioned in Section 5: Ancient History (1-8), Middle Ages (9-16), Contemporary history (17-24) and WWII plus Cold War (25-32) The results are reported in Table 6. The gender difference is very persistent and stable across periods, with an average values of the gender dummy never lower than 2. The regressions reported in Table 6 show other interesting facts. Ancient History was played better by students that specialized in Classic high school as

Table 4: Mean Size Statistics by Gender

	(1)
d-bootstrap	0.553***
	(3.76)
g-bootstrap	0.552***
	(3.76)
N	245

t statistics in parentheses

d is Cohen's (d) with standard error bootstrapped with 200 repetitions

g Hedges's (g) with standard error bootstrapped with 200 repetitions

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

well as Scientific high school. This should not come as a surprise, since the *Licei* schools specialize in subjects such as ancient latin and philosophy, while the Classic curriculum focuses also on ancient greek.

6.2 Is the gender difference driven by violent events?

One possible explanation of the gender difference may be linked to a male tendency to record more easily than females violent events. As we mentioned in section 2, the psychological literature refers to this phenomenon as to "the male warrior hypothesis" (McDonald et al. 2012). Beyond the biological tendency, the male knowledge of violent episodes may also be related to the specialization of boys in games and activities that deal with historical violent episodes (violent video games, violent movies, etc.). To tackle this issue, we create two different sub-games, one that include only violent events and the second the includes only non-violent ones. The distinction between the two is far from obvious, but we did construct different samples. "Violent events" include wars and assassinations. Less obvious is the position of an event like the French Revolution that involved some violence and assassinations, but represents also a philosophical and political revolution. The last column of Table 1 reports the classification that we use in this section.

Results are reported in Table 7. Regressions are performed using 12 violent events, in order to have the same number of items in the two sub-games. We try the same regressions with all the number of violent events and results are unchanged. Coherently with the "male warrior hypothesis" applied to history, the gender difference is stronger in the sub-game that include only violent events, but it is stil sizable also among non violent events. The fall in the gender difference is in the order of 20 percent, since the coefficient β falls from -3.8 in the column (4) to -3.1 in column (2). Overall, the results show that male subjects do have a tendency to do better in positioning violent events, but this fact can account for no more than 20 percent of the overall difference.

Table 5: Gender differences through the Timeline

	(1) Score in Time Line Test	(2) Score in Time Line Test	(3) Score in Time Line Test
Female	-25.34***	-26.19***	-22.80***
	(-4.06)	(-4.26)	(-3.71)
High School grade on History		13.40***	10.25**
		(3.37)	(2.55)
Final High School Grade		0.373	0.541*
		(1.31)	(1.93)
Classic High School			14.66*
			(1.70)
Scientific High School			9.237
			(1.24)
Parents Education			0.786
			(1.38)
Number of Books			3.055
			(1.25)
Foreign Students			-5.642
			(-0.49)
anno_eco			3.635
			(0.92)
_cons	50.85***	-87.08***	-118.4***
	(11.43)	(-2.79)	(-3.96)
N	245	244	241
R^2	0.062	0.142	0.218

t statistics in parentheses

Standard Errors robust to heteroskedasticity.

In the third column we included all the variables that we asked in the questionnaires.

Parents Education is the sum of father and mather education.

Classic and Scientific High School are what in Italy is called Liceo

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Gender bias in sub periods

	(1)	(2)	(3)	(4)
VARIABLES	score_ancient	score_Middle Age	score_Contemporary	score_ColdWar
Female	-2.567***	-2.261*	-3.726***	-2.031*
	(0.898)	(1.176)	(1.091)	(1.114)
High School grade on History	0.254	2.490***	0.834	2.056***
	(0.584)	(0.753)	(0.766)	(0.769)
Final High School Grade	0.060	0.032	0.123**	0.057
	(0.045)	(0.058)	(0.057)	(0.058)
Classic High School	5.145***	3.144*	2.420*	1.314
	(1.387)	(1.778)	(1.400)	(1.885)
Scientific High School	2.181*	1.221	1.050	0.852
	(1.155)	(1.384)	(1.337)	(1.337)
Parents Education	0.146	0.155	0.158	0.113
	(0.107)	(0.105)	(0.114)	(0.116)
Number of Books	0.043	0.360	0.614	-0.120
	(0.315)	(0.385)	(0.382)	(0.357)
Foreign Students	1.067	-2.481	-0.234	-0.183
	(1.570)	(2.359)	(1.995)	(2.261)
Constant	80.782***	57.045***	68.562***	68.981***
	(5.285)	(6.026)	(5.651)	(5.805)
Observations	240	240	240	240
R-squared	0.146	0.164	0.172	0.097

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 7: Gender bias in non violent and violent sub games

	(1)	(2)	(3)	(4)
VARIABLES	score_nonv	score_nonv	score_viol	score_viol
Female	-3.601***	-3.185**	-4.739***	-3.840**
	(1.371)	(1.297)	(1.710)	(1.716)
High School grade on History		0.745		4.094***
		(0.876)		(1.012)
Final High School Grade		0.124**		0.018
		(0.062)		(0.075)
Classic High School		5.330**		3.287
		(2.103)		(2.479)
Scientific High School		2.218		1.881
		(1.742)		(2.146)
Parents Education		0.180		0.172
		(0.166)		(0.177)
Number of Books		0.299		0.577
		(0.429)		(0.645)
Foreign Students		-1.564		-0.795
		(2.926)		(3.019)
Constant	90.591***	67.111***	89.066***	46.840***
	(0.961)	(8.459)	(1.273)	(9.257)
Observations	244	240	244	240
R-squared	0.027	0.129	0.029	0.149

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

7 Conclusion and further research

The historical sense of time is a subtle concept, and Western societies invest time and resources to create adults with such sensibility. Yet, it is traditional very difficult to understand whether this basic skill has been acquired by students after 13 years of study. Standardized tests do not focus on measuring the historical sense of time, even though this paper argues that it belongs to history as much as simple algebra belongs to math. The paper proposes a simple test for measuring quantitatively a human sense of history without relying on any specific mnemonic date. The *linetime* test is proposed for history, but can easily applied to other disciplines such as literature and philosophy. Providing such tests for other humanities is part of our current research agenda. In this sense, this paper helps developing quantitative measures of skills in a variety of humanities.

The test was administered in an experimental setting to 250 university Italian students in economics. The results show that the average students wrongly place each of the 32 events provided by at least two positions. Nevertheless, we know that undergraduate students are generally capable to place historical events that occur at the tails of the time line. In other words, events such as the Egyptian pyramids and the fall of the Berlin Wall are correctly positioned at the far end of the perceived timeline. Yet, events that occurred in the Middle Age, such as the fall of the Byzantine empire or the 100 Years War between France and England appear largely obscure to the average student. Replicating the test in other countries is important, we will pursue this research agenda in the future.

Existing literature on differences in cognitive abilities across genders suggests that female subjects outperform male subjects in memory related tests. Perhaps surprisingly, the experimental results of our *linetime test* feature a sizable difference across genders that survive various robustness checks. Our research shows that female subjects perform on average below male subjects by as much as 50 percent of the standard deviation of the entire sample and of the male subjects. The gender difference is very robust in our experiment. With respect to the existing vast knowledge on gender differences in cognitive abilities (Halpern, 2012), the paper uncovers a new place where such differences arise. Of course such a sizable difference calls for explanation. The paper investigates the male warrior hypothesis, and finds that the sizable difference is reduced by no more than 20 percent. Future research on this difference is certainly needed.

References

[1] Benbow , C. P. , Stanley , J. C. (1980). Sex differences in mathematical ability: Fact or artifact? *Science*, 210 , 1262 1264 . doi:10.1126/science.7434028

[2] Corriere della Sera (2016) *Sorpresa negli studi: donne piú brave negli studi ma penalizzate nel mercato del lavoro.* Availabe on line on www.corriere.it

[3] Davies, I. (2010). *Debates in history teaching.* London: Routhledge.

[4] Grondin S. (2010). Timing and time perception: a review of recent behavioral and neuroscience findings and theoretical directions. *Attention, Perception and Psychophysics*, April, Volume 72, Issue 3, pp 561?582

[5] Halpern, D. (2004) A Cognitive-Process Taxonomy for Sex Differences in Cognitive Abilities, *Current Direction in Psychological Science*, Vol. 13(4): 135-139

[6] Halpern D. (2012). Sex Differences in Cognitive Abilities. *Psychology Press*, New York, Fourth Edition.

[7] Hartmann, U. and Hasselhorn, M. (2008). Taking a standardized measure for an aspect of students historical thinking. *Learning and Individual Difference*, 18(2), 264-270.

[8] Hedges , L. V. , Nowell , A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals . *Science* , 269 , 41 45 . doi:10.1126/ science.7604277

[9] Huijen T., Van Boxtel C., Van de Grift W., Holthuis P. (2014).Testing elementary and secondary school students ability to perform historical perspective taking: the constructing of valid and reliable measure instruments. *European Journal of Psychology Education.*

[10] HuffPost, (2011) How Much Students Know About History?, by Davine Ravitch , available onhttps://www.huffingtonpost.com/diane-ravitch/

[11] Kautz, T. Heckman, James J. , Diris R, Bas ter Weel, Lex Borghans (2017) *Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success* NBER Working Paper No. 20749

[12] Mc Donald, Melissa, Navarrete Carlos David, and Mark Van Vugt (2012) Evolution and psychology of intergroup conflict: the male warrior hypothesis, *Philosophical Transaction of the Royal Society* 367-670-679

[13] Mayer R. (2006) Learning to teach you young people how to think historically: a case study of one student teacher's experience. *The Social Studies*, (March-April) 69-76

[14] Niederle M., Vesterlung L. (2010). Explaining the gender gap in math test scores: the role of competition. *Journal of economic perspectives.* Volume 24, Number 2, Pages 129-144.

[15] OECD (1999) Measuring Students Knowledge and Skills, *Organization for Economic Cooperation and Development.* Paris

[16] OECD (2017) Pis 2015 Results. Student's financial literacy. *Organization for Economic Cooperation and Development.* Paris

[17] VanSledright, B. A. 2004). What does it mean to think historically and how do you teach it? *Social Education* 68 (3): 230-33

[18] VanSledright, B. A. (2012), *The Challange of Rethinking History Education, Routledge.* London.

Figure 1: The Theoretical Distribution of $LT\_SCORE(J)$

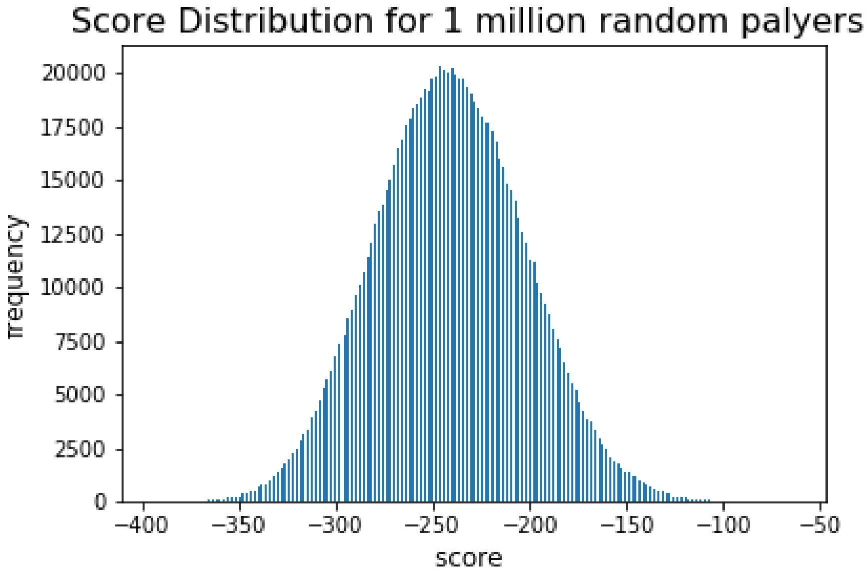


Figure 2: Distribution of the linetime test

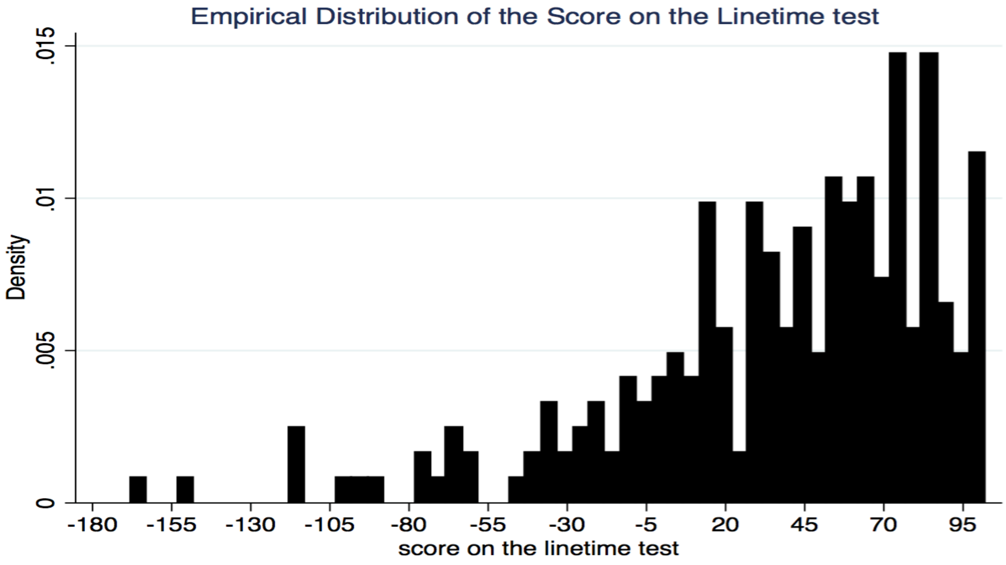


Figure 3: The Average Error in the Linetime test
See Table 1 for labeling of events

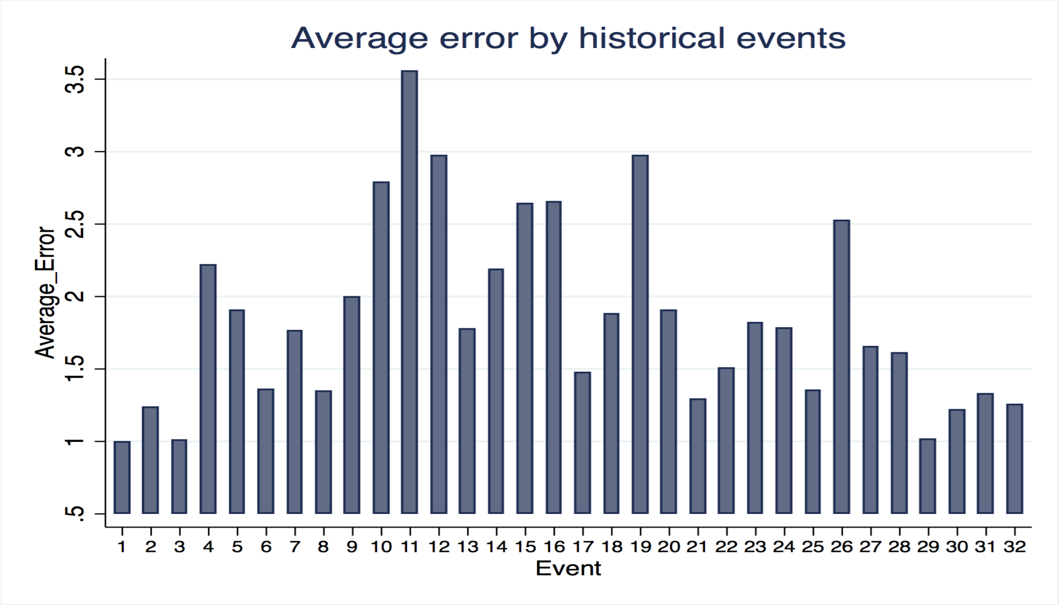
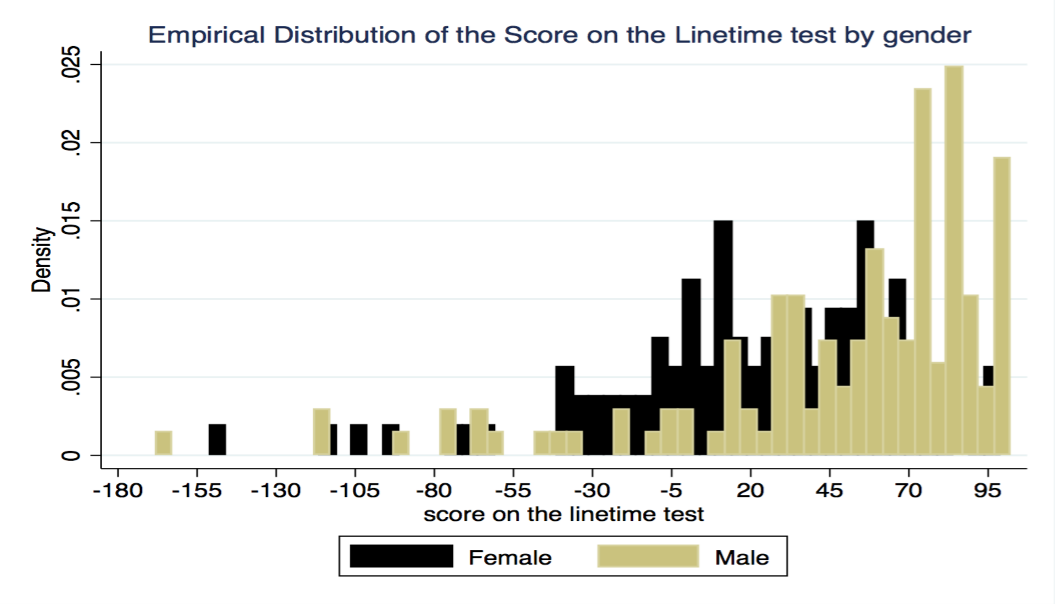


Figure 4: Distribution of the linetime test by gender


Empirical Distribution of the Score on the Linetime test by gender
Density
score on the linetime test
Female
Male