Gender-Neutral Language and Gender Disparities[ŧ]

Alma Cohen,[*] Tzur Karelitz,[**] Tamar Kricheli-Katz,[***]

Sephi Pumpian,[**] and Tali Regev[****]

This study investigates empirically whether and how the use of gender-neutral language affects the performance of women and men in real high-stakes exams. We make use of a natural experiment in which the institute administering Israel's standardized college admission tests amended the language used in its exams, making test language more gender neutral. We find that the change to a more gender-neutral language was associated with a significant improvement in the performance of women on quantitative questions, which meaningfully reduced the gender gap between male and female performance on these questions. However, the change did not affect female performance on verbal questions nor male performance on either quantitative or verbal questions. Our findings are consistent with the hypothesis that gendered language may introduce a "stereotype threat" that adversely affects women's performance in tasks in which they are stereotypically perceived to underperform. Our findings have significant implications for the ongoing academic and policy discussions regarding the use and effects of gender-neutral language.

JEL Classification: D83, I20, I24, J16, Z13

[ŧ] We thank Ian Ayers, Lucian Bebchuk, Daniel Bird, Raquel Fernandez, Steven Lehrer, Adriana Robinson, Noray Savannah and Analia Schlosser, for valuable comments and discussion. We also want to thank all the participants of the reason and decision retreat at Tel-Aviv University and all the participants at the ISPA (Israeli Psychometric Association) conference.

[*] Harvard Law School, the Eitan Berglas School of Economics, Tel Aviv University, NBER, CEPR, and ECGI.

[**] Israeli National Institute of Testing and Evaluations (NITE).

[***] The Buchman Faculty of Law, Tel Aviv University.

[****] Tiomkin Economics School, Reichman University.

"If we spoke a different language, we would perceive a somewhat different world."
Ludwig Wittgenstein[1]

I. Introduction

This paper makes use of a natural experiment to investigate empirically whether and in what way the use of gender-neutral language affects certain performance outcomes. We find that, in standardized high-stakes exams, the use of a more gender-neutral language improves women's performance. The improvement takes place in certain tasks where there is traditionally a gender gap in performance between men and women.

Languages vary in whether and how they encode gender. Even in languages that are more gender neutral, like English, some parts of speech deviate from gender neutrality by signaling that the prototypical individual is a man (for example, prototypical police officers and firefighters have traditionally been referred to as policemen and firemen, and "he" was traditionally used as the pronoun for a generic person). In recent years, however, there has been a substantial movement toward using more gender-neutral language. Thus, for example, with respect to official communications and documents, in 2021 the U.S. House of Representatives adopted rules requiring the use of gender-neutral language in House of Representatives communications. Several U.S. states, including California and New York, now require the use of gender-neutral language in all official documents and forms, already in 1987, the United Nations adopted guidelines for using such language in its official documents and communications. In contrast, after adopting similar rules in 2015, the French government reversed them in 2022, taking the position that the masculine is a neutral form that should be used in official documents for terms applicable to both women and men.

Education is a major area in which policies promoting gender-neutral language have been adopted or considered (see, e.g., National Council of Teachers of English, 2018). Of particular interest is the decision of the U.S. Educational Testing Service, which administers the SAT (Scholastic Aptitude Test) and plays a key role in U.S. college admissions, not to mandate the use of gender-neutral language in examinations after having considered making such a change (Educational Testing Service, 2022).

[1] *Remarks on Colour* (Oxford: Blackwell, 1977, ed. G. E. M. Anscombe and trans. Linda Schättle).

The policies discussed above are likely to be motivated at least in part by a belief that using gender-neutral language positively affects performance and outcomes. Therefore, it is worthwhile to obtain empirical evidence on whether, and in what settings and with what measures, making language gender neutral does indeed have such effects.

This research provides such empirical evidence. By using evidence from a natural experiment to address identification issues, we are able to identify a causal link between using a more gender-neutral language and improving women's performance in certain tasks in real-world standardized tests. Although there is significant empirical literature on the subject that uses a cross-sectional approach or a laboratory experimental approach, our study is, to the best of our knowledge, the first to provide natural-experiment evidence on this topic.

The natural experiment we used was conducted by Israel's National Institute of Testing and Evaluation (NITE), which administers the Psychometric Entrance Test (PET). PET is an SAT-like standardized test that is used for admissions to Israeli universities. We show that the transition from addressing test takers in the singular masculine to addressing them in the plural masculine, which is a more gender-neutral form, positively affected the performance of women without adversely affecting the performance of men. This use of more gender-neutral language had a positive, economically meaningful effect on the performance of women in quantitative questions. The change increased women's success by 1.5 percentage points in quantitative questions, on average. The size of this effect was about one-fifth of the original gender gap between the performance of men and women in quantitative questions. In contrast, the change in language had no effect on women's performance on verbal questions nor on men's performance on quantitative or verbal questions.

Our findings are consistent with the "stereotype threat" mechanism that has been documented in various settings. The large body of literature on the stereotype threat has shown that when gender stereotypes are evoked (sometimes merely by making gender more salient), people behave in a manner reflecting them (Bracha & Cohen, 2018; Spencer et al., 1999; Steele & Aronson, 1995). Because women are stereotypically viewed as worse than men in math, making gender more salient in a setting in which math tasks need to be performed can lead to poorer performance by women by increasing their anxiety and cognitive load or decreasing their levels of effort and attention. Consistent with the stereotype threat mechanism, we find that when women are addressed in a form that does not activate gender stereotypes, they perform better on math questions (but not on verbal questions).

While our study is the first to provide real-world causal evidence from a natural experiment on the issue of gendered address, we rely on a growing body of literature on the effects of the grammatical features of languages on people's behavior. A significant part of this literature uses cross-country studies to examine associations between linguistic features, such as grammatical structures, and the behavior of speakers (Ayres et al., 2023; Chen, 2013; Galor et al., 2020; Mavisakalyan et al., 2018; Robert et al., 2015). For example, studies using cross-country variation have identified correlations between gendered languages and gender inequality in the labor force (Gay et al., 2013; Prewitt-Freilino et al., 2012; Shoham & Lee, 2018) or gender gaps in the level of education (e.g., Davis & Reynolds, 2018; Galor et al., 2020; Jakiela & Ozier, 2018). It is widely understood, however, that despite the richness and value of cross-country studies, there are limitations on the degree of causality that can be inferred from them, due to problems such as omitted variables bias and simultaneity.

Another significant set of empirical studies has pursued an experimental approach. These studies have examined how the performance of participants in the lab was affected by variations in the linguistic features of the text presented to them. For example, such studies examined the association between gendered language and on the expression of sexist attitudes (e.g., Wasserman & Weseley, 2009), motivation (e.g., Vainapel et al., 2015), and, most relevant to our setting, performance in math tasks (Kricheli-Katz & Regev, 2021a, b, with results consistent with ours). Whereas experimental studies are not afflicted by some of the identification issues involved in cross-country studies, questions arise regarding the extent to which experimental findings can predict outcomes in real-world settings.[2]

Finally, and most broadly, our analysis is related to a large body of literature in linguistics and philosophy regarding the relationship between language and behavior (Ladd et al., 2018). Whereas some universalist linguists view the different languages people use as sharing deep-seated structures (e.g., Chomsky,1957), other linguists who hold the linguistic relativity view (Everett, 2013; Levinson, 2012; Whorf, 1956) argue that the linguistic formats that tend to vary across languages shape our perceptions and behavior.

Before proceeding, we would like to note that in the natural experiment we analyze, the change to more gender-neutral language also made the questions more inclusive of non-binary identities. Therefore, the effects of inclusiveness and the effects of neutrality cannot be

[2] In particular, outcomes in an experiment might be influenced by participants' recognition that they are taking part in an experiment, and lab experiments usually cannot fully simulate real-world settings.

disentangled. The remainder of the paper proceeds as follows. In Section II, we provide the relevant linguistic and institutional background and describe the natural experiment that enabled us to test for causal effects. In Section III, we provide our empirical analysis, and in Section IV we present our conclusions.

II. Institutional Background and the Natural Experiment

A. Gender-Neutral and Non-Gender-Neutral Texts

Gender-neutral language refers to a person in a format that does not reveal the person's gender. Standard uses of most languages[3] have long included elements that are not gender neutral. In some languages (grammatical gender languages), such as German, Romance languages, Arabic, and Hindu, every noun has a grammatical gender. For example, in such languages, the word for a woman student and a man student would not be the same. In other languages (natural gender languages), such as English, Danish, and Swedish, while nouns are mostly gender neutral, personal pronouns are specific to the particular gender. Thus, in English, an exam instruction such as "the student should open the blue book" is gender neutral, but an exam instruction stating that "the student should open his blue book" would indicate that the text has a man student in mind. In such a case, to make the instruction gender neutral, it could be changed to "the student should open their blue book" or "the student should open his or her blue book."

In the psychometric tests used in our natural experiment, the language of the test is Hebrew. Hebrew, like German, is a grammatical gender language in which nouns generally have a gender assigned to them and the noun's gender affects the form of the verb used with it and the form of the pronoun used to refer to it. For our context, it is relevant that in Hebrew verbs are also associated with gender, and thus, the verb takes a different form depending on whether a man or a woman is, for example, asked to write or to answer something.[4]

For many years prior to the change examined in this paper, Israel's NITE used exams that employed the *singular* masculine form of verbs in its instructions to test takers, thereby signaling that the writers of the text viewed men as the prototypical test takers. When making

[3] Exceptions include Estonian, Finnish, and Hungarian, none of which have either a grammatical gender or gender-specific personal pronouns.

[4] For example: the command to *write* is spelled and pronounce ktov for a man and kitvi for a woman, and the command *answer* is spelled and pronounced anne for a man, and ani for a woman.

the change, NITE switched to using the *plural* masculine form of the verb,[5] which is understood to refer more inclusively to both men and women.[6] In colloquial Hebrew there is hardly any use of the female plural, and the male plural is perceived to refer to both men and women. We refer to such a format for addressing a test taker as more gender neutral.

Hebrew has some modal verbs that are pronounced differently depending on the gender of the person being addressed by them but are spelled the same for both genders.[7] Because PET test takers received PET instructions in written form, they could read such modal verbs as addressed to them regardless of gender. Thus, for instructions that used such terms, the form of address was gender-neutral both before the switch to the plural form and after the switch. We refer to these modal verbs as *gender-ambiguous.*

B. The PET

Many countries use a standardized test for university admissions. For example, the two tests used in the United States are the SAT and the American College Test (ACT). As mentioned, Israel's comparable test, the PET, is administered by Israel's NITE.

The PET serves as an important component of the admissions process for institutions of higher education in Israel, and, like the SAT, it is designed to measure cognitive abilities, mathematical reasoning, and verbal skills. The test is administered four times a year in many locations around the country. It is available in various languages, but a substantial majority of test takers sit for the Hebrew version, and our focus is on these test takers.

The test consists of three domains: quantitative reasoning, verbal reasoning, and English proficiency. There are two sections in each of the three domains. In addition, there are two pilot sections, which are similar to the other sections but are included only for score calibration, quality assurance, and testing new questions for future use. These sections are not scored as part of the official test but are structured so that test takers are not aware that they are "pilot" sections. Therefore, test takers must treat all sections with the same degree of seriousness.

[5] *Write* in plural masculine is spelled and pronounced kitvu, and *answer* in the plural masculine is spelled and pronounced anu.

[6] The Hebrew Language Academy: https://hebrew-academy.org.il/2010/10/04/איך-פונים-לקבוצה-שרובה-נשים/

[7] For example, *you must* is spelled the same in Hebrew for both men and women but pronounced differently: alecha for a man and alayich for a woman.

The quantitative section contains 25 questions that test various areas of mathematics, such as geometry, algebra, percentages, averages, ratio questions, drawing conclusions from a diagram, and more. The mathematical knowledge required for the quantitative reasoning sections is comparable to the lowest level of mathematics required for Israel's high school matriculation exam.

The verbal section contains 30 questions that cover analogies, logic, and inferences as well as reading comprehension questions.

C. *The Natural Experiment*

In December 2009, Israel's NITE changed the form of address used in the PET from the singular masculine to the plural masculine to create a more gender-neutral environment for all test takers. Using this change as a natural experiment enables us to compare test takers' performance in a real-life setting before and after the change. To account for potential confounders, we focus on a number of sections given before and after the change, where no change was made in the content of the questions themselves but only in the form of address. By focusing on these sections, we are able to compare test takers' performance before and after the change for identical questions.

The change affected some questions while leaving other questions unaffected. Thus, by comparing differences in performance between questions that were and were not affected, we are able to control for additional confounding effects that have occurred over time.

There are two types of questions that were affected by the change to more gender-neutral language. The first type includes questions that were previously addressed in the singular masculine form and were changed to the plural masculine form (we refer to them as *gendered address* questions). The second type includes questions that were previously addressed in a singular masculine form but use a verb which is spelled in a gender-ambiguous way and were changed to the plural masculine (we refer to them as *gender ambiguous* questions).

Based on the literature mentioned above, we predict that the change from a singular- to a plural-masculine form of address will improve women's performance on quantitative questions. More specifically, we expect to see improvement with the *gendered address* questions only. This is because, unlike the plural-masculine form and the gender-ambiguous form, the singular masculine has the potential to activate a stereotype threat for women in tasks in which they are stereotypically perceived to underperform by making gender more salient and by creating a feeling of exclusion among the women.

We also do not expect to see any effect from the change to gender-neutral language on women's performance on verbal questions. This is because women are not stereotypically perceived to be worse than men on these questions, and therefore the stereotype threat is not expected to be activated. We also do not expect to find an effect on men's performance on the *gendered address* or the *gender ambiguous* questions, regardless of the type of question, (quantitative or verbal). The reason is that there is no substantial difference between the singular- and the plural-masculine forms of address for men test takers, as both forms address men in the masculine gender.

III. Analysis

A. Data and summary statistics

We obtained data on all first-time test takers who took the exam sometime between 2000 and 2012 and answered one of the repeated sections in the PET. We limit our analysis to first-time test takers, as people retaking the exam are more likely to ignore instructions because they are already familiar with them. Including people retaking the exam in our analysis potentially could have led to an understatement of the effect of the change to gender-neutral language.

We regard a section as a *repeated section* if there were no more than three questions that were replaced the second time it was administered. We exclude any altered questions from our analysis. During our sample period (2000–2012), there were 9 quantitative repeated sections (in one of them only one question was replaced), and 24 verbal repeated sections (in four of them only one question was replaced, and in 10 of them, three questions were replaced).

In each section there are three types of questions: *gendered address* questions, *gender ambiguous address* questions, and questions that have no reference to gender (*non-gendered* questions). Some questions were connected through common instructions, such as consecutive questions referring to the same graph. We omit these questions from our main analysis because we cannot know if and to what extent test takers might refer back to the instructions. We include these questions in our robustness analysis.

On average, there were two *gendered address* questions in each of the nine repeated quantitative sections (12% of questions), and on average 3.11 *gendered address* questions in each of the 24 repeated verbal sections (18% of questions). About 11% of the quantitative questions and 10% of verbal questions were *gender ambiguous* questions.

Our sample includes data from all 154,265 first-time test takers who took the Hebrew version of one of the repeated sections (quantitative or verbal) during our sample period (2000–2012). Of these, 45,082 took one of the repeated quantitative sections and 109,183 took one of the repeated verbal sections. About two-thirds of the test takers in our sample took the PET after the change to a more gender-neutral language.

Table 1 presents summary statistics for the test takers who were tested before and after the change to gender-neutral language. The data contains information on 18,909 (26,173) test takers who took one of the quantitative (verbal) sections before the change and 26,173 (73,264) test takers who took one of the quantitative (verbal) sections after the change. More women took the test (55%, which fits the official data); however, there are no significant differences between genders in participation before and after the change, or in the type of section (quantitative or verbal). Relatedly, there are no significant differences in test takers' ages or incomes. Nonetheless, test takers who took the PET after the change to gender-neutral language tended to have more educated parents, which can be explained by an increase in higher education over the years for the entire population, and by the share of immigrants in the population with a higher education (mainly those coming from the former Soviet Union). For robustness purposes, we replicate our analyses using only Israeli-born test takers.

Table 2 presents the success rate (the success of answering a question correctly) by gender, time (before and after), and type of question (quantitative and verbal), based on 2,524,334 questions completed by test takers. The average success rate in quantitative questions increases for women from 59.5% before the change to gender-neutral language to 63.2% after the change, and for men from 68.7% to 70.7%. The gender gap in the period before and after the change remained similar, at around 8%.

The improvement in the verbal questions was less substantial (from 65.6% to 66.2% for women, and 67.8% to 68.6% for men), with a negligible gender gap.

B. Empirical Strategy

We study the relationship between the form of address and test takers' performance by running the following OLS regression model:

$$
\begin{aligned}
y_{iqct} = {} & \beta_1 Gendered_q + \beta_2 Gendered_q X After_t + \beta_3 Gendered_q\ X\ Female_i \\
& + \beta_4 Gendered_q\ X\ After_t X\ Female_i + \beta_4 X_{iq} + \gamma_c + \mu_q + \delta_i + +\varepsilon_{iqt}
\end{aligned}
$$

In this regression, y_{iqct} is a binary indicator of whether person i answered question q correctly in section c, given that the section was taken at time $t.$ $Gendered$ is a dummy variable which is equal to 1 if the question included a singular masculine address before the change and plural masculine address after the change and 0 otherwise, $After$ is a dummy variable equal to 1 if the repeated section was given in the period after the policy change and 0 otherwise.

$Gendered\ X\ After$ is the interaction between the dummy specifying whether the question is a *gendered address* question and whether the repeated section was given in the period after the policy change. This interaction variable captures changes that happened over time. $Gendered\ X\ Female$ is the interaction between the dummy specifying whether the question is a *gendered address* question and $Female$, which is a dummy variable equal to 1 if the test taker is a woman, and 0 otherwise. This interaction variable captures whether the success rate for the specific question was different for men and women. $Gendered\ X\ After\ X\ Female$ captures the three-way interaction between whether the question is gendered, whether it was taken in the period after the policy change, and whether the test taker is a woman. This three-way interaction variable, which is our main variable of interest, captures whether the change in the form of address had an especially large effect on women test takers.

We also included $Question\ No.$ (the question placement within the section) to control for fatigue and $Question\ No. X\ Female$, which is an interaction between the question placement and whether the test taker is a woman, and is meant to allow for different fatigue levels between women and men. In the quantitative questions, we also control for whether the question concerned graphs, geometry, or other (as the default).

In all models, we control for the section and test takers' fixed effects. The section fixed effect captures differences between the various sections, while the test takers' fixed effect captures any difference between the different test takers and enable us to conduct within test-taker analysis, estimating the relative improvement of test takers in questions in which the form of address was gendered compared to questions which have no reference to gender.

C. Main Specification Results

Table 3 presents our main results. Columns (1)–(3) present the results of our main specification. As mentioned above, we omit connected questions from our main analysis because we do not know if and to what extent test takers refer back to the instructions for connected questions. Table 3 column (1) presents the results for the quantitative questions for all test takers, both men and women. The coefficient of the interaction $Gendered\ X\ After$ is close to zero, suggesting that there was no difference in performance on gendered questions relative to non-gendered questions before and after the change. Column (1) also suggests that the gendered questions were more difficult for women test takers. The coefficient of the interaction $Gendered\ X\ Female$ is negative and equal to -1% and statistically significant at the 5% level. The three-way interaction coefficient $Gendered\ X\ After\ X\ Female$, which is our variable of interest, is positive and statistically significant at the 5% level, suggesting that after the policy change, women's success increased by 1.5 percentage points on average in quantitative questions with a gendered address (relative to quantitative questions without a gendered address). This represents 2.4% of the 61.6% mean success rate of women in quantitative questions. To better understand the magnitude of this effect, recall that the gender gap was about 8% (see Table 2), and thus, the effect of the switch to gender-neutral language reduced the gender gap by about 20%.

Columns (2) and (3) show the results of the model when it is run separately for women and men. Column (2) indicates that gendered questions were more difficult for women than were non-gendered questions. However, there was an improvement in performance in these questions after the change in the policy, with an effect similar to that obtained in column (1), with a significance level of 1%. As for men, column (3) shows that men did better in these questions than in non-gendered questions, but we see no effect of the change from singular masculine to plural masculine for men.

In columns (4) to (6) we add to our main specification information about another type of question, the *gender ambiguous* questions, which we indicate with the variable $Ambiguous$, and its interactions with $After$ and $Female$. As noted earlier, some forms of gendered address are spelled the same way for the singular male and singular female but pronounced differently. We therefore hypothesize that because women could interpret this form of address in the singular feminine form before the policy change, the move to plural-masculine would have a smaller effect or no effect at all (as there would be no indication of the gender of the test taker in either case, there would be no activation of the stereotype threat for women).

Column (4) shows that the coefficient of $Ambiguous$ is negative and statistically significant at the 1% level, suggesting that these questions were more difficult than the other questions for both women and men test takers. The interaction $Ambiguous\ X\ Female$ is positive and statistically significant at the 1% level, suggesting that women perform better in these questions relative to non-ambiguous questions. However, the three-way interaction $Ambiguous\ X\ After\ X\ Female$ is small in magnitude and not significant. It is interesting to note that although gender-ambiguous questions are on average harder, the interaction between gender-ambiguous and women is positive and statistically significant, at a magnitude similar to what we obtain in column (1).. This suggests that women perform better when addressed in the more gender neutral, masculine plural form, even when the questions are more difficult. The fact that we do not see any difference between the periods before and after the change is consistent with the assumption that women test takers perceive the *gender ambiguous* form of address and the plural masculine form of address as more gender neutral. As before, columns (5) and (6) provide the results for the specification that includes the *gender ambiguous* questions separately for women and men test takers.

D. Robustness Tests

Table 4 provides the results of some robustness tests that we performed. Table 4, column (1) presents the results of adding a more demanding specification to our main model. In this specification, instead of controlling only for section fixed effects, we also control for $Section\ X\ Question\ No.$ fixed effect. The coefficient of our variable of interest (which is the three-way interaction $Gendered\ X\ After\ X\ Female$) remains significant and similar in magnitude (1.3% and statistically significant at the 5% level).

In column (2) we exclude the *gender ambiguous* questions. Again, we find that the coefficient of our main variable of interest, the three-way interaction, is similar in magnitude and statistically significant, which means that our results are robust to the inclusion or exclusion of these questions.

Next, as shown in columns (3) and (4), we test whether our results are robust to the inclusion of questions that we characterized as "connected" to a *gendered address* question and excluded from our main specification. In column (3), we add these questions and code them as non-gendered questions, and in column (4), we add these questions but code them as gendered questions. In both cases, we would expect the coefficient of our main variable of interest to be weaker. Indeed, we find that when the connected questions are added, and regardless of how they are coded, the effect of the change on women’s performance is smaller in magnitude than in our main specification. In both columns, the effect is negative and statistically significant at the 5% level, with an effect of -1.2% when the questions are coded as non-gendered and -0.9% when the questions are coded as gendered.

To rule out the possibility that our results were obtained by chance, we conducted a placebo test. We randomly selected two questions from the quantitative section and defined them as “$Placebo\ Gendered\ Address$” questions. We then ran our main specification (Table 3, column (1)) using the “$Placebo\ Gendered\ Address$” indicator instead of the “$Gendered\ Address$” indicator.

We repeated this procedure 1,000 times, obtaining 1000 coefficients for the three-way interaction “$Placebo\ Gendered\ Address\ X\ After\ X\ Female.$” The distribution of these 1000 coefficients is presented in Figure 1. The probability of obtaining a coefficient larger than 0.015 was found to be less than 10%.

Table 5 presents the results of this model when it is applied to verbal questions. The findings indicate that the change did not have a statistically significant effect on the success of either women or men in verbal questions.

IV. Conclusion

Our study investigates the effect of using more gender-neutral language on the performance of women and men in high-stakes standardized exams. To this end, we have taken advantage of a natural experiment that enabled us to identify whether gender-neutral language is causally linked to changes in performance.

We find that using a more gender-neutral language improved the performance of women on quantitative questions in the standardized test we considered. The effect was both statistically significant and economically meaningful, with a magnitude roughly equal to one-fifth of the gender disparity between men’s and women’s scores on such questions. Our findings suggest that using language that is not gender neutral exacerbates the gender gap between men and women by introducing a stereotype threat, and that a change to gender-neutral language can reduce this gender gap by weakening the stereotype threat with minimal costs.

Note that our within-test-taker specification can only detect the effect of the change on gendered questions relative to non-gendered questions. But, if the overall performance of women in the test improved due to the policy change, we are not able to capture this positive change in our analysis. In that case, our results underestimate the true effect of the policy change.

Our results have significant implications.

Among other things, they suggest that the organizations administering the SAT and ACT standardized college tests in the United States could reconsider their long-standing position of including non-gender-neutral language in their test questions, perhaps by conducting trials about the issue, such as the change carried out by Israel's NITE and analyzed in this paper. Beyond standardized tests, our findings suggest that policies supporting gender-neutral language, which have been increasingly debated and implemented, could well have practical effects on gender disparities in behavior and outcomes. Most broadly, our paper contributes to the heated debate about the influence that the structure of languages has on human behavior, going back to classic theorists such as Chomsky (1957), Sapir (1951) and Whorf (1956). Our findings are consistent with and support the views regarding the inextricable links between language structures and human behavior.

References

Ayres, Ian, Tamar Kricheli-Katz, and Tali Regev (2023), “Do Languages Generate Future-Oriented Economic Behavior?.” *Proceedings of the National Academy of Science.*

Bracha Anat, Alma Cohen and Lynn Conell-Price (2013), “The Heterogeneous Effect of Affirmative Action on Performance,” *Journal of Economics Behavior and Organization*, Vol. 158, pp. 172-218.

Carroll, Mary, Christiane Von Stutterheim, and Ralf Nüse.(2004). "The language and thought debate: a psycholinguistic approach." In Multidisciplinary Approaches to Language Production, ed. T Pechmann, C Habel, pp. 183–218. Berlin: Mouton de Gruyter.

Chen, Keith M. (2013), “The Effect of Language on Economic Behavior: Evidence from Savings Rates, Health Behaviors, and Retirement Assets.” *American Economic Review*, Vol 103, No. 2, pp. 690–731.

Chomsky, Noam. (1957). *Syntactic Structures*, The Hague: Mouton.

Educational Testing Service (2022), ETS Guidelines for Fair Tests and Communications, https://www.ets.org/content/dam/ets-org/pdfs/about/fair-tests-and-communications.pdf.

Everett, Caleb. (2013) "Linguistic relativity." In Linguistic Relativity. De Gruyter Mouton.

Galor, Oded, Ömer Özak, and Assaf Sarid, (2020), "Linguistic traits and human capital formation." in *AEA Papers and Proceedings*, Vol. 110, pp. 309-13.

Gay, Victor, Estefania Santacreu-Vasut and Amir Shoham (2013), "The grammatical origins of gender roles," In: Working papers Berkeley Economic History Laboratory Paper Series.

Jakiela, Pamela and Owen Ozier (2018), “Gendered Language.” Policy Research working paper; no. WPS 8464. Washington, D.C.: World Bank Group. http://documents.worldbank.org/curated/en/405621528167411253/Gendered-language

Kricheli-Katz, Tamar, and Tali Regev (2021a), “The effect of language on performance: Do gendered languages fail women in maths?,” *NPJ science of learning*, Vol. 6, no. 1, pp. 1–7.

Kricheli-Katz, Tamar, and Tali Regev (2021b), “Does the Hebrew Language fail women? Results from five experiments.” *Israeli Sociology*, pp. 80–100.

Ladd, D. Robert, Seán G. Roberts, and Dan Dediu (2015), "Correlational studies in typological and historical linguistics," *Annu. Rev. Linguist.* Vol 1.1, pp. 221–241.

Levinson, Stephen C. (2012) Foreword. In Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf, ed. JB Carroll, SC Levinson, P Lee, pp. vii–xxiii. Cambridge, MA: MIT Press. 2nd ed.

Lewis Davis, and Megan Reynolds (2018), "Gendered language and the educational gender gap." *Economics Letters*, Vol. 168, pp. 46–48.

Mavisakalyan, Astghik, Yashar Tarverdi and Clas Weber (2018), "Talking in the present, caring for the future: Language and environment." *Journal of Comparative Economics*, Vol. 46(4), pp: 1370–1387.

National Council of Teachers of English (2018), *Statement on Gender and Language*.

Sapir, Edward (1951[1929]), "The Status of Linguistics as a Science," in *Selected Writings*, David Mandelbaum, ed. Berkeley: University of California Press. Orig. pub. in Language 5, pp: 207-214.

Spencer, S. J., Steele, C. M. & Quinn, D. M. (1999), "Stereotype threat and women's math performance," *J. Exp. Soc. Psychol*, Vol. 35, pp. 4–28.

Prewitt-Freilino, Jennifer L., Andrew T. Caswell and Emmi K. Laakso (2012), "The gendering of language: A comparison of gender equality in countries with gendered, natural gender, and genderless languages." *Sex Roles*, Vol. 66, pp. 268 –281.

Shoham Amir, and Sang Mook Lee (2018), "The Causal Impact of Grammatical Gender Making on Gender Wage Inequality and County Income Inequality," *Business and Society*, Vol. 56, No. 6, pp. 1216–1251

Steele, C. M. & Aronson, J. (1995), "Stereotype threat and the intellectual test performance of African Americans," *J. Pers. Soc. Psychol.*, Vol. 69, pp. 797–811.

Whorf, Benjamin. Carroll, John B. (ed.) (1956), *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*, MIT Press.

Figures

Figure 1: Placebo test results of randomly assigning a "*Placebo Gendered Address*."

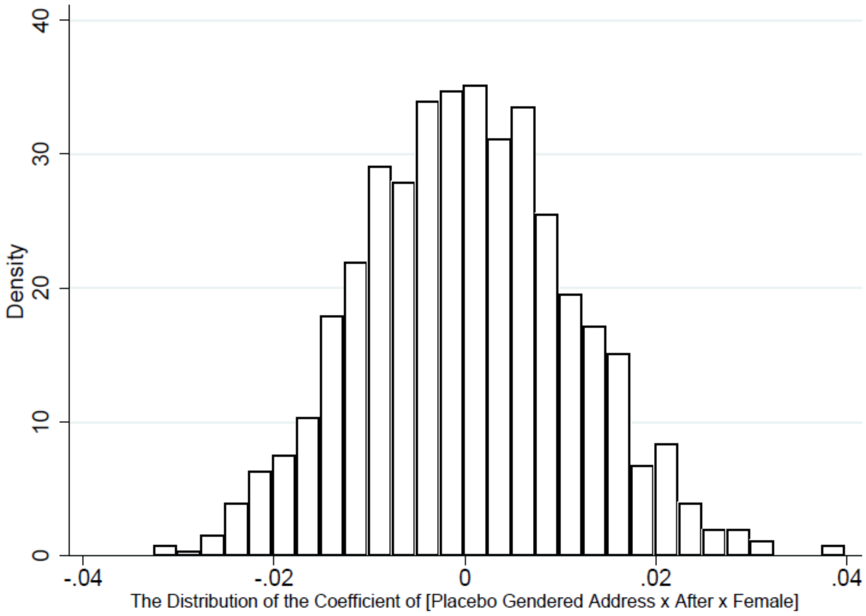
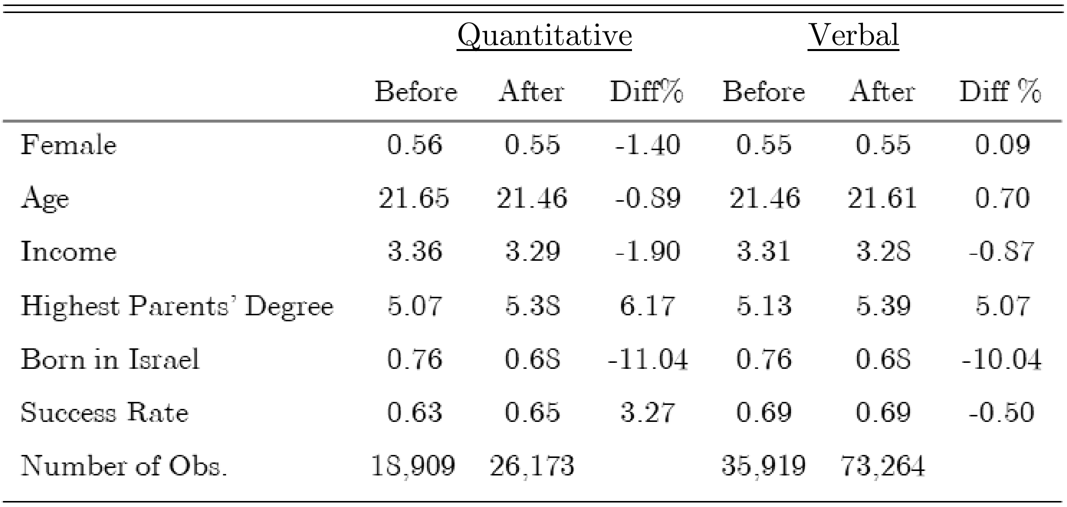


Table 1: Descriptive Statistics – per Test Takers by Exam Domain Type (Quantitative vs. Verbal), Before and After the Change



	Quantitative			Verbal		
	Before	After	Diff%	Before	After	Diff %
Female	0.56	0.55	-1.40	0.55	0.55	0.09
Age	21.65	21.46	-0.89	21.46	21.61	0.70
Income	3.36	3.29	-1.90	3.31	3.28	-0.87
Highest Parents' Degree	5.07	5.38	6.17	5.13	5.39	5.07
Born in Israel	0.76	0.68	-11.04	0.76	0.68	-10.04
Success Rate	0.63	0.65	3.27	0.69	0.69	-0.50
Number of Obs.	18,909	26,173		35,919	73,264	

Table 2: Descriptive Statistics – Success Rate for Type of Questions, Gender, and Before and After the Change

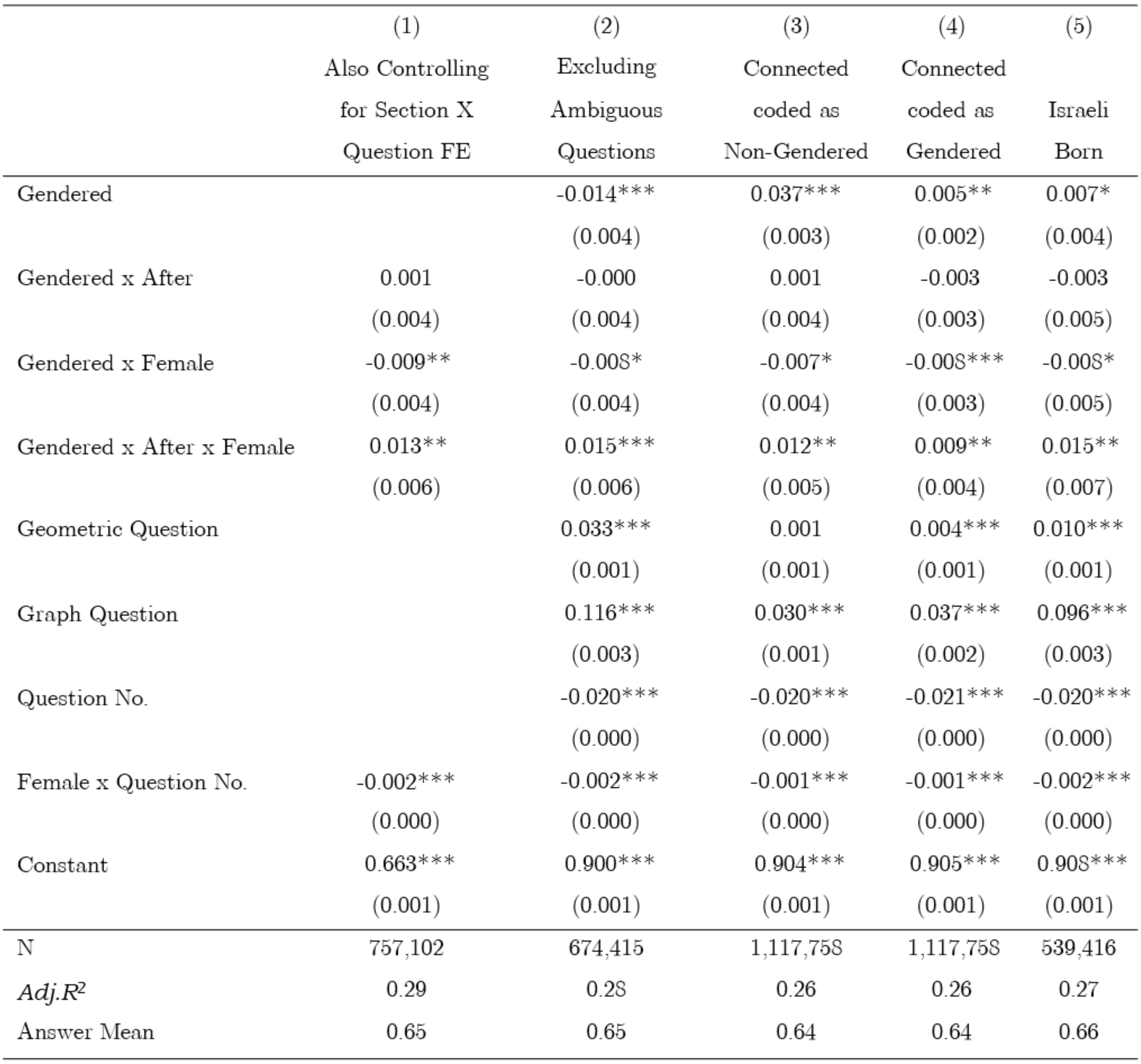
Type Of Question	Gender	Before	N	After	N	Diff %
Quantitative	Women	0.595	179,242	0.632	242,676	6.27
	Men	0.687	139,517	0.707	195,667	2.80
Verbal	Women	0.656	319,661	0.662	650,077	0.89
	Men	0.678	263,621	0.686	533,873	1.14

Table 3: Question Success Rate and Form of Address (quantitative questions)

	(1)	(2)	(3)	(4)	(5)	(6)
	All	Women	Men	All	Women	Men
Gendered	0.000	-0.015***	0.007*	-0.012***	-0.026***	-0.004
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Gendered x After	-0.000	0.015***	-0.001	-0.001	0.015***	-0.001
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Gendered x Female	-0.010**			-0.008*		
	(0.004)			(0.004)		
Gendered x After x Female	0.015**			0.015***		
	(0.006)			(0.006)		
Geometric Question	0.014***	0.028***	-0.002	0.030***	0.043***	0.013***
	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)
Graph Question	0.105***	0.113***	0.094***	0.113***	0.122***	0.103***
	(0.003)	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)
Question No.	-0.020***	-0.022***	-0.020***	-0.020***	-0.022***	-0.020***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Female x Question No.	-0.002***			-0.001***		
	(0.000)			(0.000)		
Ambiguous				-0.045***	-0.036***	-0.036***
				(0.004)	(0.004)	(0.004)
Ambiguous x After				-0.009*	-0.005	-0.007
				(0.005)	(0.004)	(0.005)
Ambiguous x Female				0.017***		
				(0.005)		
Ambiguous x After x Female				0.005		
				(0.006)		
Constant	0.897***	0.866***	0.936***	0.900***	0.869***	0.939***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
N	757102	421918	335184	757102	421918	335184
$Adj.R^2$	0.27	0.27	0.27	0.27	0.27	0.27
Answer Mean	0.65	0.62	0.70	0.65	0.62	0.70

Note: We also control for Section and Test-Taker ID Fixed Effects. Standard errors are in parentheses and are clustered by Test-Taker ID. Stars denote the level of statistical significance $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 4: Robustness Tests



	(1) Also Controlling for Section X Question FE	(2) Excluding Ambiguous Questions	(3) Connected coded as Non-Gendered	(4) Connected coded as Gendered	(5) Israeli Born
Gendered		-0.014***	0.037***	0.005**	0.007*
		(0.004)	(0.003)	(0.002)	(0.004)
Gendered x After	0.001	-0.000	0.001	-0.003	-0.003
	(0.004)	(0.004)	(0.004)	(0.003)	(0.005)
Gendered x Female	-0.009**	-0.008*	-0.007*	-0.008***	-0.008*
	(0.004)	(0.004)	(0.004)	(0.003)	(0.005)
Gendered x After x Female	0.013**	0.015***	0.012**	0.009**	0.015**
	(0.006)	(0.006)	(0.005)	(0.004)	(0.007)
Geometric Question		0.033***	0.001	0.004***	0.010***
		(0.001)	(0.001)	(0.001)	(0.001)
Graph Question		0.116***	0.030***	0.037***	0.096***
		(0.003)	(0.001)	(0.002)	(0.003)
Question No.		-0.020***	-0.020***	-0.021***	-0.020***
		(0.000)	(0.000)	(0.000)	(0.000)
Female x Question No.	-0.002***	-0.002***	-0.001***	-0.001***	-0.002***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Constant	0.663***	0.900***	0.904***	0.905***	0.908***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N	757,102	674,415	1,117,758	1,117,758	539,416
Adj.R^2	0.29	0.28	0.26	0.26	0.27
Answer Mean	0.65	0.65	0.64	0.64	0.66

Note: We also control for Section and Test-Taker ID Fixed Effects. Standard errors are in parentheses and are clustered by Test-Taker ID. Stars denote the level of statistical significance $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p <$ 0.01.

Table 5: Success Rate and Form of Address (Verbal Questions)

	(1) Verbal-all	(2) Verbal-women	(3) Verbal-men
Gendered	0.022***	0.013***	0.022***
	(0.002)	(0.002)	(0.002)
Gendered x After	-0.000	-0.000	-0.000
	(0.002)	(0.002)	(0.002)
Gendered x Female	-0.009***		
	(0.003)		
Gendered x After x Female	-0.000		
	(0.003)		
Question No.	-0.010***	-0.011***	-0.010***
	(0.000)	(0.000)	(0.000)
Female x Question No.	-0.001***		
	(0.000)		
Constant	0.850***	0.850***	0.850***
	(0.001)	(0.001)	(0.002)
N	1,767,232	969,738	797,494
Adjusted R-Squared	0.19	0.18	0.18
Answer Mean	0.67	0.66	0.68

Note: We also control for Section and Test-Taker ID Fixed Effects. Standard errors are in parentheses and are clustered by Test-Taker ID. Stars denote the level of statistical significance $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.