Gender-Neutral Language and Gender Disparities

Alma Cohen,[[1]](#footnote-1)\* Tzur Karelitz,[[2]](#footnote-2)\*\* Tamar Kricheli-Katz,[[3]](#footnote-3)\*\*\*

Sephi Pumpian,\*\* and Tali Regev[[4]](#footnote-4)\*\*\*\*

In this study, we present an empirical examination of whether and how the use of gender-neutral language affects the performance of women and men in real high-stakes exams. We take advantage of a natural experiment in which the institute administering Israel’s standardized college admission tests amended the language used in its test papers to make it more gender-neutral. We find that this change was associated with a significant increase in the performance of women in quantitative questions, which meaningfully reduced the gender gap between the performance of men and women in these questions. By contrast, the change did not affect the performance of women in verbal questions nor the performance of men in either quantitative or verbal questions. Our findings are consistent with the hypothesis that language evoking gender 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.

1. Introduction

This paper uses a natural experiment to investigate empirically whether and in what way the use of gender-neutral language affects behavior. We find that in standardized high-stakes exams, the use of gender-neutral language improves the performance of women at certain tasks where there is 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 person is male (for example, prototypical police officers and firefighters were referred to as policemen and firemen, and “he” was used as the pronoun for a generic person). In recent years, however, there is substantial support for and movement toward using more gender-neutral language. Thus, for example, with respect to official communications and documents, the US House of Representatives adopted rules requiring the use of gender-neutral language in House of Representatives communications, several US states including California and New York required the use of gender-neutral language in all official documents and forms, and the United Nations adopted guidelines for using such language in its official documents and communications. By contrast, after adopting such 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 in favor of gender-neutral language have been adopted or considered (see, e.g., National Council of Teachers of English (2018)). For example, the Educational Testing Service, which administers the SAT (Scholastic Aptitude Test), which plays a key role in US college admissions, has considered mandating the use of gender-neutral language in questionnaires but decided not to do so (Educational Testing Service (2022)).

The policies discussed above are likely to be at least partially motivated by a belief that using gender-neutral language affects behavior and outcomes. Therefore, it is worthwhile to obtain empirical evidence on whether, and in what settings and ways, making language gender-neutral has such effects.

We provide such empirical evidence. By using evidence from a natural experiment to address identification issues, we are able to identify a causal link between using gender-neutral language and improving the performance of women 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 the subject.

The natural experiment we use was carried out by the Israeli 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 effect that is economically meaningful on the performance of women in quantitative questions. The change increased women’s success by, on average, 1.5 percentage points in quantitative questions. The size of this effect was about one-fifth of the original gender gap between the performance of men and women in quantitative questions. By contrast, the change had no effect on the performance of women in verbal questions nor on the performance of men in 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 according to them (Steele and Aronson (1995)), Spencer et al (1999), Bracha and Cohen (2018)). Because women are viewed as less good at math, making gender more salient in a setting in which math tasks need to be performed can lead to a worse performance by women by increasing their anxiety and cognitive load or decreasing their levels of effort and attention. Consistent with a stereotype threat channel, we find that women perform better on math questions (but not on verbal questions) when they are not addressed in a form that activates gender stereotypes by assuming that men are the ”prototypical test takers.”

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 and grammatical structures and the behavior of speakers (Ayres et al (2023), Chen (2013), Mavisakalyan et al (2018), Galor et al (2020), Robert et al (2015)). For example, studies using cross-country variation have identified correlations between gendered languages and gender inequality in the labor force (Prewitt-Freilino et al. (2012), Gay et al. (2013), Shoham and Lee (2018)) or gender gaps in the level of education (e.g. Davis and Reynolds (2018), Jakiela and Ozier (2018), Galor et al (2020)). 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 pursues an experimental approach. These studies examine how the performance of participants in the lab is affected by variations in the linguistic features of the text presented to them. For example, such studies examine the association between gendered language and sexist attitudes (e.g. Wasserman & Weseley (2009)), between gendered language and motivation (e.g. Vainapel et al. (2015)), and, most relevant to our setting, the association between gendered language and performance in math tasks (Kricheli-Katz and Regev (2021a,b), which finds 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.[[5]](#footnote-5)

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 (Whorf (1956), Levinson (2012), Everett (2013)) 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 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 enables us to test for causal effects. In Section III, we provide our empirical analysis, and in Section IV we present our conclusions.

1. Institutional Background and the Natural Experiment
2. *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. In most languages,[[6]](#footnote-6) standard uses of language have long had elements that were not gender-neutral. In some languages (Grammatical Gendered Languages), such as German, Romance languages, Arabic, and Hindu, every noun has a grammatical gender; in such languages for example, the term for a female student and a male student would not be the same. In other languages (Natural Gender Languages), such as English, Danish, and Swedish, while personal nouns are mostly gender-neutral, personal pronouns are specific to the particular gender. Thus, in English, a test instruction such as “the student should open the bluebook” is gender-neutral, but a test instruction stating that “the student should open his bluebook” would indicate that the text has a male student in mind. In such a case, to make the instruction gender-neutral, it could be changed to “the student should open their bluebook” or “the student should open his or her bluebook.”

In the tests used in our natural experiment, the questions are written in Hebrew. Hebrew is similar to German in that it is a Grammatical Gendered Language in which nouns generally have a gender assigned to them and the gender of a noun 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 verbs are also associated with gender, and thus when a man is asked, say, to write or to answer, the verb has a form that is different than when a woman is asked to write or to answer.[[7]](#footnote-7)

For many years prior to the change examined in this paper, NITE used questionnaires that instructed test-takers using the masculine form of verbs, which signaled that the writers of the ­­text had men as the prototypical test-takers. When making the change, NITE switched to using the plural masculine form of the verb, which is understood to refer to both men and women.[[8]](#footnote-8) We refer to such a format for addressing a test-taker as 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.[[9]](#footnote-9) 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 *unisex*.

1. *The PET*

Many countries use a standardized test for university admission. For example, the two tests used in the US are the SAT and the ACT. Israel similarly has such a test, which is referred to as PET and is administered by the Israeli NITE.

The PET serves as an important component of the admissions process for institutions of higher education in Israel. The test, which is similar in nature to the SAT test, is designed to measure cognitive abilities, mathematical reasoning, and verbal skills. The test is administered in many locations around the country, four times a year. 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 sections: mathematical reasoning, verbal reasoning (including a writing assignment), and proficiency in English. There are two chapters in each of the three sections. In addition, there are two pilot chapters, which are similar to the other chapters but are only included for score calibration, quality assurance, and testing new questions for future use. These chapters are not scored as part of the official test but are structured so that test-takers are not aware that they are “pilot” chapters. Therefore, test-takers have to treat all chapters with the same degree of seriousness.

The quantitative chapter contains 20 questions that cover problems in various areas of mathematics such as geometry, algebra, percentages, averages, ratio questions, drawing conclusions from a diagram, etc. The mathematical knowledge required for the quantitative chapters is comparable to the lowest level of mathematics required for the high-school matriculation exam.

The verbal chapter contains about 25 questions that include analogies, logic, and inference questions, as well as reading comprehension questions.

1. *The Natural Experiment*

In December 2009, the Israeli 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. We use this change as a natural experiment, which allows 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 chapters given before and after the change, where no change has been made in the content of the questions themselves but only in the form of address. By focusing on these chapters, 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 gender-neutral language. The first type includes questions that were previously addressed in the singular masculine form and were changed to the plural masculine (we refer to them as *gendered address* questions). The second type includes questions that were previously addressed in the unisex form and were changed to the plural masculine (we refer to them as *unisex* questions).

Based on the literature mentioned above, we predicted that the change from a singular to a plural-masculine form of address would improve women’s performance on quantitative questions. More specifically, we expected to see the improvement only for the *gendered address* questions. This is because, unlike the plural-masculine form and the unisex form, the singular masculine has the potential to activate the stereotype threat for women in tasks in which they are known to underperform, by making gender more salient and by excluding them.

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

1. Analysis

*A. Data and summary statistics*

We obtained data on all first-time test-takers who took the exam between 2000 and 2012 and answered one of the repeated chapters in the NITE. We limited 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 may, therefore, have led to an understatement of the effect of the change to gender-neutral language.

We regard a chapter as a *repeated chapter* 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 8 quantitative repeated chapters (in one of them there was only one question that was replaced), and 24 verbal repeated chapters (in four of them only one question was replaced, and in 10 of them three questions were replaced).

We define a question to be a *gendered address* question if it was addressed in the singular masculine form before the change. We define a question to be a *unisex address* question if it was addressed in the unisex form before the change. Some questions are connected through common instructions, such as consecutive questions referring to the same graph. We drop these questions from our main analysis because we do not 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 are two *gendered address* questions in each of the nine repeated quantitative chapters (12% of questions), and on average 3.11 *gendered address* questions in each of the 24 repeated verbal chapters (18% of questions). About 11% of quantitative questions and 10% of verbal questions are *unisex address* questions.

Our sample includes data from all 154,265 first-time test-takers who took the Hebrew version of one of the repeated chapters (quantitative or verbal) during our sample period (2000–2012). Of these, 45,082 took one of the repeated quantitative chapters and 109,183 took one of the repeated verbal chapters. About two-thirds of the test-takers in our sample took the test after the change to 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) chapters before the change and 26,173 (73,264) test-takers who took one of the quantitative (verbal) chapters after the change. More females 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 chapter (quantitative or verbal). Relatedly, there are no significant differences in test-takers’ ages or incomes. Nonetheless, test takers who took the exams 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 whole population, and by the share of immigrants with higher education (mainly coming from the USSR). For robustness purposes, we replicate our analyses using only Israeli-born test-takers.

Table 2 presents the success rate 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%. From the table, we perceive that 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 studied the relationship between the form of address and test-takers’ performance by running the following OLS regression model:

In this regression, is a binary indicator of whether person answered question correctly in chapter , given the test was taken at time 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, is a dummy variable equal to 1 if the repeated chapter was given in the period after the policy-change and 0 otherwise.

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

We also included (the question placement within the chapter) to control for fatigue and , which is an interaction between the question placement and whether the test-taker is female to allow for different fatigue levels between women and men. In the quantitative questions, we also controlled for whether the question concerned graphs, geometry, or other (as the default).

In all models, we controlled for the chapter and test-takers’ fixed effects. The chapter fixed effect captures differences between the various chapters, while the test-takers fixed effect captures any difference between the different test-takers and enabled us to conduct an analysis by test-taker, estimating the relative improvement of test-takers in questions where the form of address was non-gendered compared to gendered questions.

*C. Main Specification Results*

Table 3 presents our main results. Column (1)-(3) presents the results of our main specification. As mentioned above, we drop 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 male and female. The coefficient of the interaction 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 shows that the gendered questions were more difficult for female test-takers. The coefficient of the interaction is negative and statistically significant at the 5% level. The three-way interaction coefficient , 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 get a sense of the magnitude of this effect, recall (see Table 2) that the gender gap was about 8%, 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, respectively. Column (2) shows that gendered questions were harder for women compared to non-gendered questions; however, there was an improvement in performance in these questions after the change in the policy, with an effect similar to what we obtained in column (1) with a significant 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 *unisex address* questions, which we indicate with the variable , and its interactions with and . 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 hypothesized that since before the policy-change women could interpret this form of address in the singular feminine form, 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).

Column (4) shows that the coefficient of is negative and statistically significant at the 1% level, suggesting that these questions were harder than non-unisex questions for both female and male test-takers. The interaction is positive and statistically significant at the 1% level, suggesting that women perform better in these questions relative to non-unisex questions. However, the three-way interaction is small in magnitude and not significant. It is interesting to note that, although unisex questions are on average harder, the interaction between unisex and female is positive and statistically significant at a similar magnitude to what we obtain in column (1) for improvement by female test-takers in gendered questions after the change to gender-neutral language. This suggests that women perform better when addressed in the feminine form, even when the questions are harder. The fact that we do not see any difference between the periods before and after the change is consistent with the fact that female test-takers perceive the *unisex* form of address and the plural masculine form of address as gender-neutral. As before, columns (5) and (6) provide the results for the specification that includes the *unisex* questions separately for female and male 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 only controlling for chapter fixed effect we also controlled for fixed effect. The coefficient of our variable of interest (which is the three-way interaction ) remains significant and similar in magnitude (1.3% and statistically significant at the 5% level).

In column (2) we excluded the *unisex* questions. Again, we found 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 tested 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 to gender-neutral language 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 questions. We then ran our main specification (Table 3, column (1)) using the indicator instead of the indicator. We repeated this procedure 1,000 times, obtaining 1000 coefficients for the three-way interaction . 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 to gender-neutral language did not have a statistically significant effect on the success of either women or men in verbal questions.

1. 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 gender-neutral language improved the performance of women in quantitative questions in the standardized exam we considered. The effect was not only statistically significant but also 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 non-gender-neutral language exacerbates the gender gap between men and women by introducing a stereotype threat, and a switch to gender-neutral language can decrease this gender gap by weakening the stereotype threat.

Our results have significant implications. Among other things, they suggest that the organizations administering the SAT and ACT standardized college tests should reconsider their long-standing position of including non-gender-neutral language in their test questions, and at a minimum should conduct experiments such as the one carried out by the NITE and analyzed in this paper. Beyond standardized tests, our findings suggest that policies supporting gender-neutral language, which have been increasingly implemented and debated, could well have practical effects on gender disparities in behavior and outcomes. Most broadly, our findings are consistent with and support the large body of literature, going back to classic theorists such as Whorf (1956) and Wittgenstein (1953), 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*.

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*.

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 - XXX.

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

Wittgenstein, Ludwig, (1953), *Philosophical Investigations*.

Figures

Figure 1: Placebo test results of randomly assigning a 'placebo gendered address'



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



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



Table3: Question Success Rate and form of Address (quantitative questions)



Table4: Robustness Tests



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



1. \* Harvard Law School, Tel-Aviv University Berglas School of Economics, NBER, CEPR, and ECGI. [↑](#footnote-ref-1)
2. \*\* Israeli National Institute of Testing and Evaluations (NITE). [↑](#footnote-ref-2)
3. \*\*\* Tel Aviv University Buchman School of Law. [↑](#footnote-ref-3)
4. \*\*\*\* Reichman University Tiomkin School of Economics. [↑](#footnote-ref-4)
5. In particular, outcomes in an experiment might be influenced by participants’ recognition that they are taking part in in an experiment, and lab experiments usually cannot fully simulate real-world settings. [↑](#footnote-ref-5)
6. Exceptions include Estonian, Finnish and Hungarian that all have neither a grammatical gender nor gender-specific personal pronouns. [↑](#footnote-ref-6)
7. For example: *write* is spelled and pronounce K’tov for a man and Kitvi for a woman, and *answer* is spelled and pronounced Ane for a man , and Ani for a woman. [↑](#footnote-ref-7)
8. *Write* in plural masculine is spelled and pronounced Kitvu, and *answer* in the plural masculine is spelled and pronounced Anu. [↑](#footnote-ref-8)
9. 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. [↑](#footnote-ref-9)