The Common Problem of Bad Controls in Tests of the Linguistic Savings Hypothesis

A Comment on Ayres et al. (*PNAS*, 2023) and related literature

Paul Clist *

Journal of Comments and Replications in Economics, Volume 4, 2025-9, DOI: 10.18718/81781.48

JEL: D14, E21, Z13
Keywords: Linguistic savings hypothesis, linguistic relativity, Sapir–Whorf hypothesis, Language, Patience

Data Availability: The Stata code and the data to reproduce the results of this replication can be downloaded at JCRE's data archive (DOI: 10.15456/j1.2025224.1531306214).

Please Cite As: Clist, P. (2025). The Common Problem of Bad Controls in Tests of the Linguistic Savings Hypothesis. A Comment on Ayres et al. (*PNAS*, 2023) and related literature. *Journal of Comments and Replications in Economics*, Vol.4 (2025-9). DOI: 10.18718/81781.48

Abstract

Languages encode the future differently, such that languages can be sorted into different types. When making future predictions, Strong Future Time Reference (FTR) languages often require marking future-time, whereas weak FTR languages do not. The Linguistic Savings Hypothesis predicts this difference will affect people's patience, perhaps because weak FTR languages make the future feel closer. Three recent articles have reported significant effects on individual patience as well as firm-level income smoothing and investment efficiency. However each paper includes controls which increase bias, rather than reduce it. Due to data sharing limitations, I can only publish a reanalysis in one case; fixing the problem leaves smaller and non-significant effects. This problem does not affect all research on the Linguistic Savings Hypothesis, but the effects are not as large or widespread as previously reported.

*School of Global Development, University of East Anglia, UK. paul.clist@uea.ac.uk, https://paulclist.github.io/
I would like to thank Sean G. Roberts and Juan D. Moreno-Ternero for useful comments, along with Tamar Kricheli-Katz, Editor Bob Reed and two anonymous referees. I am not aware of any conflicts of interest.

Received August 12, 2025; Accepted November 16, 2025; Published December 18, 2025.

1 The Linguistic Savings Hypothesis

Languages describe the world in different ways. While they share common linguistic building blocks, they differ in what is obligatory. For example, in German you could say the equivalent of 'tomorrow it is raining'. In English you would need to mark the future explicitly, such as 'it *will* rain' or 'it *is going to* rain'. A distinction can then be drawn between weak and strong 'future time reference' (FTR) languages (Chen, 2013; Dahl, 2013). German is called a 'weak FTR language' because it can use the present tense to describe the future, whereas in English a future tense is often obligatory to refer to the future.[1]

The Linguistic Savings Hypothesis is that someone's time preferences will differ according to whether they speak a strong or weak FTR language (Chen, 2013). The original evidence includes higher savings rates and less risky activities (e.g. smoking) amongst speakers of weak FTR languages. This was both across countries and within countries between speakers of languages that differ, with reasonably large and significant differences in a range of samples.

This potentially has consequences for a range of important behaviour, with recent research examining possibilities at the firm-level including the use of variable pay, CSR performance, and tax avoidance (see Cao et al., 2024, for references). Since the Linguistic Savings Hypothesis was first proposed, I identify three relevant publications in high level general interest journals, comprising all research on the topic of the Linguistic Savings Hypothesis published in Management Science and the Proceedings of the National Academy of Sciences[2] (Kim et al., 2021; Cao et al., 2024; Ayres et al., 2023). All three articles suffer from the problem of bad controls, affecting their inferences and conclusions. I proceed by discussing the statistical problem.

2 Defining Bad Controls

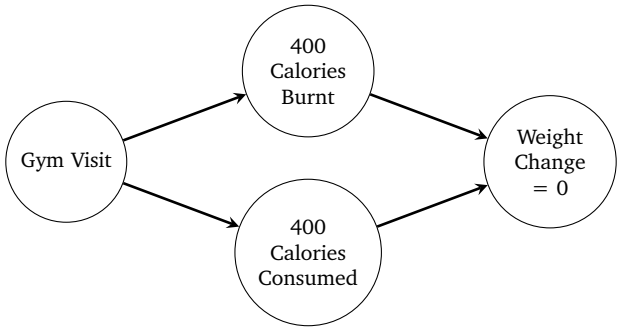
Researchers wishing to identify the effect of some X on Y must decide whether to include controls Z. Aware of the problem of omitted variable bias, they tend to include many controls, hoping to increase the precision of their estimate. Some are relatively less aware of 'bad controls', situations in which controls increase the bias of their estimate. To help researchers in this choice a common rule-of-thumb has gained some traction "Do not control for post-treatment variables" (Gelman & Hill, 2006, p. 188).

This rule of thumb is misleading; it is "neither necessary nor sufficient for deciding whether a variable is a good control" (Cinelli et al., 2022, p.3). Rather, Wooldridge (2024, p.1) provides an excellent two-sentence summary of his previous three-page paper (Wooldridge, 2005): "... I formally showed how, when a treatment assignment is randomized, unconfoundedness fails when control variables are included whose distributions differ by treatment status. Specifically, if W is the

[1] The precise distinction is subtle, and simplified in the text in the service of readability. Dahl (2013) categorised languages as Strong FTR if future predictions in the main clause, without intention, require the use of marking future-time. But English *can* use the present tense to refer to the future in some instances, such as 'The train leaves at 9pm tomorrow'. As such, the distinction is perhaps better thought of as a spectrum than a binary distinction, with Chen (2013) and others relying on Dahl (2013)'s categorisation, which captures common usage. For more on modality and usage, see Robertson & Roberts (2023) and Robertson et al. (2025).

[2] I searched for "Future Time Reference" and (separately) "Linguistic Savings Hypothesis" in Management Science, PNAS, Science and Nature for the period 2013-2024. These three papers are the only articles which ask whether FTR affects some outcome measure.

Figure 1: Playful Example of Bad Controls

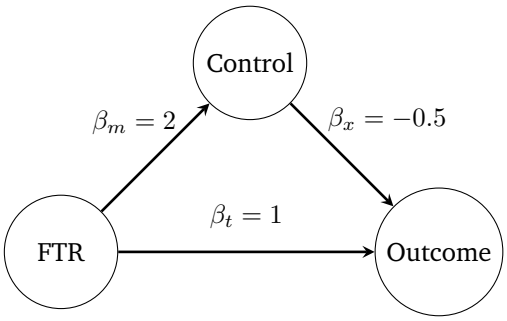


treatment variable and $\boldsymbol{X}$ is the vector of control variables, $D(\boldsymbol{X}|W) \neq D(\boldsymbol{X})$, where $D(\cdot)$ denotes distribution, provides a working definition of "bad controls" in randomized controlled trials." This definition of bad controls is also helpful for observational studies, and shows that the older advice can create problems. Some pre-treatment variables' distributions differ by treatment status, even if they were determined long before treatment. They should therefore be excluded, or they will introduce bias. The more useful, and correct, rule of thumb is therefore whether the distribution of a control variable differs by treatment status. If it does, it should not be included if we wish to estimate the total effect of X on Y.

Consider a playful example, as per figure 1. Imagine that a person's weight is a simple formula based on calories consumed minus calories burnt. You wish to see whether your gym helps you lose weight, so collect data. After every gym visit you *cannot* resist a smoothie, which has as many calories as you burn in an average gym visit. If we just think of the effect of going to the gym on weight, we would then estimate something like $weight = constant + \beta_g \cdot gym\ visits + error$ and find the total effect of visiting the gym, correctly concluding that gym visits don't help ($\hat{\beta_g} \approx \beta_g \approx 0$). A bad control might condition on the snack, estimating $weight = constant + \beta_g * \cdot gym\ visits + \beta_s \cdot snack + error$. This would no longer find the total effect of the gym visit, parcelling out the snack and exercise elements, and mistakenly lead you to increase your visits to no avail. Whilst the coefficient may be *labelled* gym visits, it would actually be capturing only the exercise part.

Alternatively, we could control for the very mechanism through which we expect the treatment to work. In this case, calories burnt. Here we would estimate $weight = constant + \beta_g \cdot gym\ visits + \beta_c \cdot calories\ burnt + error$, and conclude gym visits are the reason for putting on weight! Again, the coefficient would be labelled gym visits, but here it would be capturing only the effect of snacks. In either case, including a bad control (distribution differs by gym visits) means we do not estimate the total effect of gym visits.

Figure 2: A hypothetical DAG, illustrating bad controls



Total FTR Effect = 0

Note: FTR stands for (weak) Future Time Reference.

Table 1: A regression using simulated data, illustrating bad controls

	(1)	(2)
FTR	1.012***	0.00324
	(112.95)	(0.46)
Control	-0.504***	
	(-159.19)	
Constant	0.000411	0.00133
	(0.09)	(0.27)
N	100000	100000

Note: t statistics in parentheses. Significance at the 10, 5, and 1% levels are denoted by 1, 2, and 3 stars, respectively. The dependent variable is 'Outcome'. The simulated data adds random noise to both control and the outcome, with a mean of 0 and a standard deviation of 1.

Now let's be a little more formal. There are many kinds of bad control (see Cinelli et al., 2022, for a clear taxonomy using a DAG framework); my focus is what is variously called concomitant variable bias, included variable bias, over-control bias and inappropriate control (respectively, see Rosenbaum, 1984; McElreath, 2020; Elwert & Winship, 2014; Wysocki et al., 2022). Let us consider a stylised example: Figure 2 shows that speaking a weak FTR language has an effect on some outcome. There is also a control which is affected by FTR, which in turn affects the outcome ($\beta_m \neq 0, \beta_x \neq 0$). If we want to see the total effect of FTR on the outcome, we should not include the control in the regression: "...we want to leave all channels through which the causal effect flows 'untouched' ... Controlling for Z [Control] will block the very effect we want to estimate (the total effect of X [FTR] on Y [Outcome]), thus biasing our estimates..." (Cinelli et al., 2022, p.8). Using figure 2 to build intuition, imagine the true values are known such that $\beta_t = 1$, $\beta_m = 2$ and $\beta_x = -.5$ (see Clist, 2025, for the code). The true total effect of FTR on the outcome is $1 + 2 \times -.5 = 0$. Table 1, column (1) shows that if we run a regression of the form $Outcome = \alpha + \beta_t \cdot FTR + \beta_x \cdot Control + \epsilon$, we will estimate $\hat{\beta}_t \approx 1$, and mistakenly conclude that FTR significantly increases the outcome.

This article is designed to apply this insight, correcting a weakness in important research. Omitting the bad controls removes over-control bias - in the hypothetical example above, column (2) of Table 1 shows the results from a regression of the form $Outcome = \alpha + \beta_{FTR} \cdot FTR + \epsilon$ correctly estimates $\hat{\beta}_{FTR} \approx 0$, leading to the correct inference of the total effect of FTR. (I do not claim it removes all other biases, of course.) Whether the control was determined before or after the treatment does not affect the bias, only whether or not its distribution differs by treatment status.

This hypothetical example illustrates a case where bad controls mistakenly lead to significant results. The opposite effect is also possible, where researchers mistakenly conclude there is no significant effect. This would happen if, for example, $\beta_t = 0.1$, $\beta_m = 2$ and $\beta_x = 1$. Researchers, including the bad control, would conclude that the effect of FTR is small at 0.1, when the true total effect is $0.1 + 2 \times 1 = 2.1$. The signs on each path determine the direction of the bias. In cases of multiple bad controls, the direction of the bias will be particularly difficult to predict.

Partial effects are, of course, sometimes of interest in their own right. To return to the playful example, researchers may be interested in whether exercise affects weight. A control is bad when a partial direct effect is interpreted as though it were the direct effect. In each case we discuss, the estimated direct partial effect of FTR is interpreted as though it describes the total effect of FTR.

We now move to real examples of bad controls, starting with observational evidence.

3 *Non-Experimental Evidence, With Bad Controls*

Kim et al. (2021) ask whether FTR affects firm-level investment efficiency. They posit that weak FTR speakers will have a lower discount rate on future returns and so underinvest less, but have an ambiguous prediction on overinvestment. Their main analysis is cross-country, but they supplement this using two within-country analyses, exploiting the multilanguage environment in Switzerland, and the different countries of birth of US CEOs. They report large effects in the cross-country analysis (a 6% difference in underinvestment), and in the Swiss (17%) and US (11%) intra-country analyses. Each effect is reported significant at the 1% level.

Their main analysis has two stages. They first run a regression of lagged sales growth on capital expenditure, scaled by lagged property, plant, and equipment. If a firm is in the top quartile of residuals - essentially if they invest more than the simple model predicts - they are classified as overinvesting, the bottom quartile are classified as underinvesting, with the middle two quartiles the benchmark group. This dependent variable is then used in the main cross-country regressions, where FTR is included alongside 23 other control variables, as well as fixed effects for continent, industry and year.

If the article's arguments are correct, the results suffer from over-control bias. The authors explicitly expect two controls will be affected by FTR, and so meet the definition of a bad control. The first is *Long Term Orientation*. The authors posit FTR will affect investment efficiency because "weak FTR speakers perceive future events to be less distant and thus tend to apply a lower discount rate" (Kim et al., 2021, p.2610). By controlling for long term orientation, they then bias the estimate of the total effect of FTR. The second is financial reporting quality. The authors argue "languages may also affect investment efficiency indirectly through their effects on financial reporting quality" (Kim et al., 2021, p.2613), noting that they found this differed by FTR in a previous article (Kim et al., 2017). Kim et al. (2021, p.2614) then conclude "[t]hus, in all our regressions, we control for financial reporting quality...". The rationale appears to be that because it is important it should be controlled for. As Wooldridge (2005, p.1028) writes "... it may seem clear that one should not include in x variables that are themselves responses to treatment, but this can get lost..."

Each of the regressions in the paper include at least one of these bad controls, akin to controlling for exercise in the playful example discussed earlier. Without the underlying data, we cannot test the arguments that these control variables really do vary by FTR (and so should be excluded), nor examine other controls, nor recover an unbiased estimate. With the information available we can merely state that the presence of a bias seems very likely, but we cannot know the direction.

The second example also uses observed data. Cao et al. (2024) are interested in whether there is a difference in income smoothing for firms headquartered in countries that speak weak or strong FTR languages. Further, they hypothesise that the effect will be more pronounced if firms have longer-term relationships with major stakeholders. The analysis includes bad controls. This is clear from the text, which reports: "...the means and medians of most control variables differ between firms in weak- versus strong-FTR countries... These significant differences underscore the importance of including the control variables in our main regression model." (Cao et al., 2024, p.7).

This shows that the authors are not trying to nefariously mislead the reader - the authors (and presumably referees and editors) appear to simply be unaware of the problem of bad controls. The intuitively plausible reasoning of the authors appears to be that the control is important and therefore should be taken into consideration. Angrist & Pischke (2009, p.64) frame the problem of bad controls using an experiment, which may help develop the intuition: "[b]ad controls are variables that are themselves outcome variables in the notional experiment at hand. That is, bad controls might just as well be dependent variables too." If we are interested in the total effect of FTR on income smoothing, bad controls hinder our efforts.

Whilst Management Science now requires datasets and code to be made available, the legal 'terms of use' preclude running and reporting any code other than the original without the permission of the authors. It was not forthcoming in this case, apparently due to data rights, and insufficient detail was provided in order to reconstruct the data from original sources. This means that we cannot report a result excluding bad controls.

In both cases, we have partial effects erroneously discussed as though they were total effects. The data sharing policy has improved, but not enough to recover and report an unbiased estimate.

4 Experimental Evidence, With and Without A Bad Control

I now discuss an experimental example, recovering an estimate without bad controls. Ayres et al. (2023) analyse three experiments, seeking to answer whether and why FTR affects individuals' time preferences. Their innovation is to randomise language using bilingual subjects who speak both a strong and a weak FTR language. This builds on previous work which randomised phrasing within the same language (Chen et al., 2019) or used bilinguals for one language pair (Clist & Hong, 2023). In experiment 1, Mturk subjects are randomly allocated to a language, and asked if they want $3 today, or if they'd prefer to wait a week for a larger amount. The experiment elicits the smallest amount that subjects would happily wait for, between $3.05 and $7. A smaller reservation price (e.g. $3.05) means subjects are more patient.

The analysis of experiment 1 appeared, prima facie, to avoid the problem of bad controls. A number of typographical errors in the original manuscript have since been corrected (the sign of a regressor, confusing significant at 5% with significant at 10%, and confusing significant at 1% and 10%), but the original table implied the only control that is always present, language proficiency, did not differ significantly by FTR. However, there was a typographical error in this table, as the standard deviation was called a 'standard error'. Inadvertently, this masked the large difference in language proficiency by FTR, meaning language proficiency is a bad control, and the reported

results are biased. The average proficiency score in strong FTR languages (8.11) is different from weak FTR languages (7.48) at the 1% level ($t = 9.57$, $N = 1130$, unpaired test, $p < 0.0000$). This is on a scale that runs 6-9, as all those scoring 1-5 were dropped from the data. As McElreath (2020, p.170) writes, "[i]ncluded variable bias is real. Carefully randomised experiments can be ruined just as easily as uncontrolled observational studies."

How can one's proficiency be a bad control, if it is predetermined? In this case, it is not a problem of randomisation, but rather more subtle. Treatment does not change proficiency scores but *which proficiency score is included in the regression*. People's proficiency in weak FTR languages is about 0.63 units lower. As these proficiency scores differ significantly by treatment, and patience varies by proficiency, controlling for proficiency *in the allocated language* fits the situation described in figure 2 and the rest of section 2.

Thankfully, the underlying data and code are available (Ayres et al., 2022), so the unbiased results can be estimated (with all code available; Clist, 2025). Whilst one may normally wish to cautiously investigate Conditional Randomization Tests in this case,[3] there appears to be a problem of construct validity. There are large differences in the proficiency scores by language, with those responding in English getting much higher scores (an average of 8.5 out of 9, whereas the other languages all have averages between 7.3 and 8). This difference in scores may be due to the way the proficiency test is implemented. It appears a single test was translated into all languages, where subjects complete 9 multiple choice questions. In some languages, the 'correct' answer appears ambiguous, potentially lowering the average score in that language. As a further indicator of a potential problem, self-identified natives score somewhat lower on proficiency. In a regression of being native on proficiency, $\hat{\beta} = -0.228$, $t = -3.33$ and $p = 0.001$.

To recover an estimate without over-control bias, Table 2 shows the results with and without the bad control. The original result is significant at the 1% level. Removing the source of bias leads to a total causal estimate that is smaller, and insignificant. To ground the interpretation of the size of the coefficient, the average reservation price for those in the strong FTR treatment is 3.97, compared to 3.85 for those in the weak FTR treatment. The results in table 2 closely resemble the hypothetical example discussed in table 1 and figure 2 as only one bad control is included. To compare, $\hat{\beta}_t = 0.5$, $\hat{\beta}_x = -0.44$, and $\hat{\beta_m} = 0.7$. The overall estimated effect of FTR on patience is not 0.5, but $0.5 - 0.44 \times 0.7 \approx 0.18$ (with the difference due to rounding and censored data).

[3] Caution is warranted as even advocates of such strategies warn that they can be used to p-hack, making the inference invalid. Johansson & Nordin (2022) recommend an algorithm to be used in conjunction with a preanalysis plan.

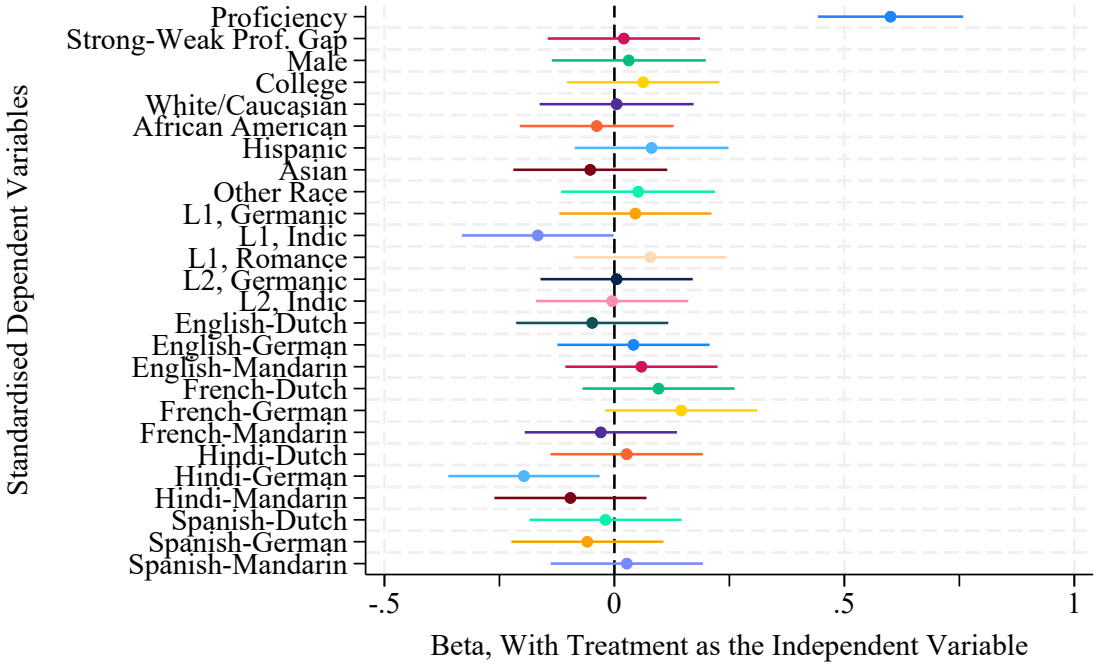
Table 2: FTR and Individual Patience, With and Without a Bad Control

Bad Controls:	Yes	No
	(1)	(2)
FTR Strong	0.50***	0.18
	(0.15)	(0.14)
Proficiency	-0.44***	
	(0.06)	
Constant	6.74***	3.49***
	(0.47)	(0.10)
Sigma	2.16***	2.38***
	(0.19)	(0.21)
Observations	523	523

Note: The dependent variable is the reservation wage, which measures impatience. A tobit model is used, censoring at 3.05 and 7. Standard Errors are provided in parentheses. Significance at the 10, 5 and 1% levels are denoted by 1, 2 and 3 stars, respectively.

Whilst reported results always include proficiency as a control, it is not the only control used. A reader may wonder whether the other controls also vary by treatment. In figure 3 I test all eight controls, plotting whether they vary by treatment status. Some have multiple levels, e.g. the control variable race has 5 levels, hence the total of 26 rows. I standardise each control to ease comparison, so the beta reports the difference in means between treated and untreated groups, in standard deviations of the control.

Figure 3: Testing Bad Controls: Effect of Treatment on Each (Standardised) Control

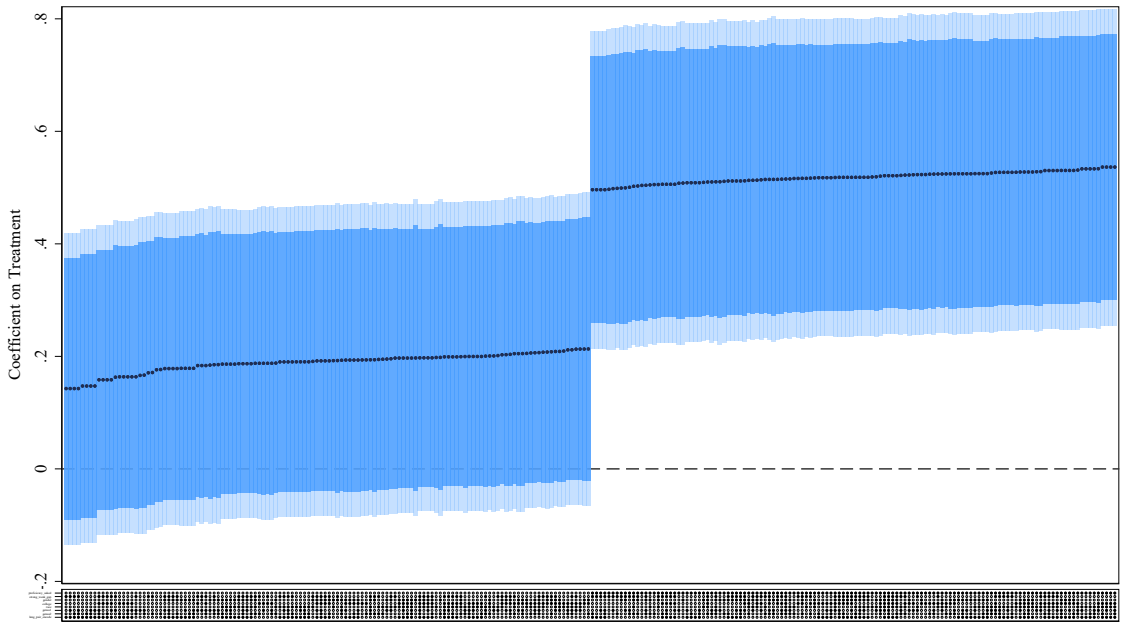


Note: In each case, I regress the treatment on the control in a bivariate regression. The figure then plots each beta with its 95% confidence interval. As per Wooldridge (2005), a bad control is defined as varying by treatment. We test for differences in means, rather than distributions, as this is easier to report. As a less powerful test, failing it provides a strong indication of problems.

We can see here that proficiency varies by 0.6 standard deviations by treatment status. No other variable is close in magnitude: it appears there is only one bad control, which unfortunately is the only control always included.

A reader may wonder whether the reported specifications are cherry picked to show unusually significant or insignificant results. Using the eight controls selected by Ayres et al. (2023), there are $2^8 = 256$ possible combinations of controls. In figure 4 I report a specification curve for all combinations, with betas ordered by size (see Simonsohn et al., 2020, for more on this method). The second panel shows which control is included, albeit with so many combinations, this is difficult to read. The important element is easy to grasp. All specifications including the bad control estimate larger effects than all specifications excluding it. Specifications including proficiency are all significant at the 5% level, with an average effect size of 0.52. Specifications excluding the bad control average 0.19, and are all insignificant at the 5% level.

Figure 4: Specification Curve: Including the Bad Control



Note: Each dot represents one beta, plotted with 90 and 95% confidence intervals. The 256 combinations are ordered by effect size. The largest half all include proficiency, the smaller half do not.

In short, the claim of a significant difference in patience by FTR hangs on including the only control which varies by treatment, i.e. the only variable which meets the definition of a bad control.

I have focused on the strategy used in Ayres et al. (2023), where covariates are added because of a lack of balance. Mutz et al. (2019) argue this practice has no basis in statistical theory, and with a sufficient number of possible control variables, can lead to any conclusion. In a comment on my original submission, the authors of Ayres et al. (2023) propose an alternative strategy of selecting a subsample on the basis of the bad control, arguing that significant effects imply we should not be concerned about overcontrol bias. Unfortunately, this is the same strategy in a different guise. This can be seen if we recall the playful example in section 2. Selecting a sample of only smoothie-drinking people will wrongly lead you to estimate a positive effect of gym-going on weight loss, as you compare gym-going-smoothie-drinkers to gym-avoiding-smoothie-drinkers. Adjusting for a bad control, or using it to select a subsample, are similar strategies that will tend to lead to biased results. Mutz et al. (2019) discuss the related idea of post-stratification (i.e. using many subsamples, and averaging over these), arguing it is largely analogous to adjusting for a bad control, and is likely to increase bias.

5 Discussion and Concluding Remarks

The issues of errors, replication and reanalysis are now more well known, with various attempts at mitigation now in place (e.g. Davis et al., 2023). We can reflect a little on progress, using these

three papers. Management Science now requires data and code to be made available, and employs a data editor to ensure the results can be reproduced (Simchi-Levi, 2020). Whilst they can be mechanically reproduced, the process did not catch several errors in how results were reported, nor catch the problem of bad controls. PNAS has a Statistical and Methodological Review Committee but no data editor. Further, whilst data should now be made available, the licence means it cannot be used for any process other than a mechanical reproduction, which places greater emphasis on catching statistical problems. In short, whilst there is progress, there appears to be some way to go.

The worry in this case is that the error of including bad controls has been normalised for a field of research. The next barrier is error correction, as Brown et al. (2018a, p.2569): argue "[n]ormalizing error correction is necessary to advance the [scientific] enterprise." Smaldino & McElreath (2016)'s argument is relevant here; error correction on its own is unlikely to be sufficient to push against the prevailing forces, such as journals' desire to publish significant results. If 'bad science' (to borrow their term) can produce significant results more efficiently than good science, bad science will tend to be rewarded over time. My hope is that this article helps readers, authors and editors to more quickly spot this particular mistake: we should only control for variables which do not vary by treatment. The problem is related to a focus on publishing significant results, as a less biased effect estimate would probably be harder to publish (Bakker et al., 2012; McShane et al., 2019).

Given the errors discussed, how should we think about the Linguistic Savings Hypothesis? Imagine a reader was interested in the topic, but had only read the top general interest journals Science, Nature, the Proceedings of the National Academy of Sciences and Management Science. They would have only seen the three articles I discuss here, which report large, significant but biased effects. Where we can recover unbiased total effects, outcomes vary by FTR but not significantly.

The last decade has of course seen other relevant research on the Linguistic Savings Hypothesis. Some work has focused on the original observational evidence, seeking to account for relatedness between languages (Roberts & Winters, 2013; Roberts et al., 2015). Other work has found FTR to significantly affect the individual time preferences of people that live in the same country but speak different languages. This is true for children in Italy (Sutter et al., 2018) and high-school students in Switzerland (Herz et al., 2021). In more applied settings, some studies use the language of migrants to try and isolate the effect of language, whilst comparing amongst people that live in the same country. There are significant effects of FTR on college attendance in the US (Galor et al., 2020), and on self-employment in Switzerland (Campo et al., 2023). Note that these studies could potentially be confounded by culture (Brown et al., 2018b), as well as an argument that historical culture affected grammar, which in turn affects preferences (Galor et al., 2017). There are also studies which show null effects (Falk et al., 2018; Clist & Hong, 2023), but I am not aware of studies which show significant effects in the other direction. It is therefore plausible that the effect is somewhat smaller and less significant than sometimes claimed, but that it does exist. There is also more fundamental research on the difference between languages that points to the role of modality, which contains information on probability, and not tense marking in FTR per se (Robertson & Roberts, 2023; Robertson et al., 2025). Future research should refine our understanding of the true effect sizes and causes in different domains. In future, bad controls should be avoided, and main effects should ideally be established without any controls, or at least without any that differ by the treatment variable.

Bibliography

Angrist, J. D. & Pischke, J.-S. (**2009**). Mostly Harmless Econometrics: An Empiricist's Companion. Princeton university press. DOI: 10.2307/j.ctvcm4j72.

Ayres, I., Katz, T. K. & Regev, T. (**2022**). "The Effects of Languages on Future Oriented Behaviors". *OSF, December 17*. DOI: 10.17605/OSF.IO/QCGPW.

Ayres, I., Katz, T. K. & Regev, T. (**2023**). "Languages and Future-oriented Economic Behavior—Experimental Evidence for Causal Effects". *Proceedings of the National Academy of Sciences*, 120(7): e2208871120. DOI: 10.1073/pnas.2208871120.

Bakker, M., Van Dijk, A. & Wicherts, J. M. (**2012**). "The Rules of the Game called Psychological Science". *Perspectives on Psychological Science*, 7(6): 543–554. DOI: 10.1177/1745691612459060.

Brown, A. W., Kaiser, K. A. & Allison, D. B. (**2018**a). "Issues with Data and Analyses: Errors, Underlying Themes, and Potential Solutions". *Proceedings of the National Academy of Sciences*, 115(11): 2563–2570. DOI: 10.1073/pnas.1708279115.

Brown, M., Henchoz, C. & Spycher, T. (**2018**b). "Culture and Financial Literacy: Evidence from a Within-country Language Border". *Journal of Economic Behavior & Organization*, 150: 62–85. DOI: 10.1016/j.jebo.2018.03.0119.

Campo, F., Nunziata, L. & Rocco, L. (**2023**). "Business is Tense: New Evidence on How Language Affects Economic Activity". *Journal of Economic Growth*, pages 1–29. DOI: 10.1007/s10887-023-09229-5.

Cao, W., Myers, L. A. & Zhang, Z. (**2024**). "The Effect of Language on Income Smoothing: Cross-Country Evidence". *Management Science*, 70(9): 5832–5852. DOI: 10.1287/mnsc.2021.01753.

Chen, J. I., He, T.-S. & Riyanto, Y. E. (**2019**). "The Effect of Language on Economic Behavior: Examining the Causal Link between Future Tense and Time Preference in the Lab". *European Economic Review*, 120: 103307. DOI: 10.1016/j.euroecorev.2019.103307.

Chen, M. K. (**2013**). "The Effect of Language on Economic Behavior: Evidence from Savings Rates, Health Behaviors, and Retirement Assets". *American Economic Review*, 103(2): 690–731. DOI: 10.1257/aer.103.2.690.

Cinelli, C., Forney, A. & Pearl, J. (**2022**). "A Crash Course in Good and Bad Controls". *Sociological Methods & Research*. DOI: 10.1177/00491241221099552.

Clist, P. (**2025**). "Replication Data for: The Common Problem of Bad Controls in Tests of the Linguistic Savings Hypothesis". *ZBW - Leibniz Informationszentrum Wirtschaft*, https://journaldata.zbw.eu/dataset/the-common-problem-of-bad-controls-in-tests-of-the-linguistic-savings-hypothesis-replication-file. DOI: 10.15456/j1.2025224.1531306214.

Clist, P. & Hong, Y.-y. (**2023**). "Do International Students Learn Foreign Preferences? The Interplay of Language, Identity and Assimilation". *Journal of Economic Psychology*, 98: 102658. DOI: 10.1016/j.joep.2023.102658.

Dahl, Ö. (**2013**). "Stuck in the Futureless Zone". *Diversity Linguistics comment*, available at https://dlc.hypotheses.org/360(09): 2013.

Davis, A. M., Flicker, B., Hyndman, K., Katok, E., Keppler, S., Leider, S., Long, X. & Tong, J. D. (**2023**). "A Replication Study of Operations Management Experiments in Management Science". *Management Science*, 69(9): 4977–4991. DOI: 10.1287/mnsc.2023.4866.

Elwert, F. & Winship, C. (**2014**). "Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable". *Annual Review of Sociology*, 40(1): 31–53. DOI: 10.1146/annurev-soc-071913-043455.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D. & Sunde, U. (**2018**). "Global Evidence on Economic Preferences". *The Quarterly Journal of Economics*, 133(4): 1645–1692. DOI: 10.1093/qje/qjy013.

Galor, O., Özak, Ö. & Sarid, A. (**2017**). "Geographical Origins and Economic Consequences of Language Structures". *Available at SSRN 2820889*. DOI: 10.2139/ssrn.2820889.

Galor, O., Özak, Ö. & Sarid, A. (**2020**). "Linguistic Traits and Human Capital Formation". *AEA Papers and Proceedings*, 110: 309–313. DOI: 10.1257/pandp.20201069.

Gelman, A. & Hill, J. (**2006**). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

Herz, H., Huber, M., Maillard-Bjedov, T. & Tyahlo, S. (**2021**). "Time Preferences across Language Groups: Evidence on Intertemporal Choices from the Swiss Language Border". *The Economic Journal*, 131(639): 2920–2954. DOI: 10.1093/ej/ueab025.

Johansson, P. & Nordin, M. (**2022**). "Inference in Experiments Conditional on Observed Imbalances in Covariates". *The American Statistician*, 76(4): 394–404. DOI: 10.1080/00031305.2022.2054859.

Kim, J., Kim, Y. & Zhou, J. (**2017**). "Languages and Earnings Management". *Journal of Accounting and Economics*, 63(2-3): 288–306. DOI: 10.1016/j.jacceco.2017.04.001.

Kim, J., Kim, Y. & Zhou, J. (**2021**). "Time Encoding in Languages and Investment Efficiency". *Management Science*, 67(4): 2609–2629. DOI: 10.1287/mnsc.2019.3555.

McElreath, R. (**2020**). Statistical Rethinking: A Bayesian Course with Examples in R and Stan, Second Edition. Chapman and Hall/CRC. DOI: 10.1201/9780429029608.

McShane, B. B., Gal, D., Gelman, A., Robert, C. & Tackett, J. L. (**2019**). "Abandon Statistical Significance". *The American Statistician*, 73(sup1): 235–245. DOI: 10.1080/00031305.2018.1527253.

Mutz, D. C., Pemantle, R. & Pham, P. (**2019**). "The Perils of Balance Testing in Experimental Design: Messy Analyses of Clean Data". *The American Statistician*, 73(1): 32–42. DOI: 10.1080/00031305.2017.1322143.

Roberts, S. & Winters, J. (**2013**). "Linguistic Diversity and Traffic Accidents: Lessons from Statistical Studies of Cultural Traits". *PloS one*, 8(8): e70902. DOI: 10.1371/journal.pone.0070902.

Roberts, S. G., Winters, J. & Chen, K. (**2015**). "Future Tense and Economic Decisions: Controlling for Cultural Evolution". *PloS one*, 10(7): e0132145. DOI: 10.1371/journal.pone.0132145.

Robertson, C. & Roberts, S. G. (**2023**). "Not When but Whether: Modality and Future Time Reference in English and Dutch". *Cognitive Science*, 47(1): e13224. DOI: 10.1111/cogs.13224.

Robertson, C., Roberts, S. G., Majid, A. & Dunbar, R. I. (**2025**). "Language and Economic Behaviour: Future Tense Use Causes less not more Temporal Discounting". *PLoS One*, 20(5): e0317422. DOI: 10.1371/journal.pone.0317422.

Rosenbaum, P. R. (**1984**). "The Consequences of Adjustment for a Concomitant Variable that has been Affected by the Treatment". *Journal of the Royal Statistical Society Series A: Statistics in Society*, 147(5): 656–666. DOI: 10.2307/2981697.

Simchi-Levi, D. (**2020**). "From the Editor: Diversity, Equity, and Inclusion in Management Science". *Management Science*, 66(9): 3802–3802. DOI: 10.1287/mnsc.2020.3759.

Simonsohn, U., Simmons, J. P. & Nelson, L. D. (**2020**). "Specification Curve Analysis". *Nature Human Behaviour*, 4(11): 1208–1214. DOI: 10.1038/s41562-020-0912-z.

Smaldino, P. E. & McElreath, R. (**2016**). "The Natural Selection of Bad Science". *Royal Society open science*, 3(9): 160384. DOI: 10.1098/rsos.160384.

Sutter, M., Angerer, S., Glätzle-Rützler, D. & Lergetporer, P. (**2018**). "Language Group Differences in Time Preferences: Evidence from Primary School Children in a Bilingual City". *European Economic Review*, 106: 21–34. DOI: 10.1016/j.euroecorev.2018.04.003.

Wooldridge, J. M. (**2005**). "Violating Ignorability of Treatment by Controlling for too many Factors". *Econometric Theory*, 21(5): 1026–1028. DOI: 10.1017/S0266466605050516.

Wooldridge, J. M. (**2024**). "A Formal Investigation of "Bad" Controls". SSRN Working Paper No. 4688295 DOI: 10.2139/ssrn.4688295.

Wysocki, A. C., Lawson, K. M. & Rhemtulla, M. (**2022**). "Statistical Control Requires Causal Justification". *Advances in Methods and Practices in Psychological Science*, 5(2): 25152459221095823. DOI: 10.1177/25152459221095823.