

# Persistence

## A critical review

**ABSTRACT:** I review, replicate and extend the analysis from seven econometric papers studying how events that happened to and values held by our ancestors affect their descendants several generations afterwards (intergenerational persistence). I argue that together the papers provide moderate evidence of the existence of long term causal effects mediated by parentage.

**KEYWORDS:** persistence, cultural persistence, economic history, multiple hypothesis testing, post design power analysis, spatial autocorrelation bias, causality, natural experiments, instrumental variables.

### HIGHLIGHTS:

- I discuss a gold standard for cultural persistence studies, covering how to (1) identify robust long term correlations via regression studies under different sets of controls, (2) identify causal effects via natural experiments and (3) identify whether culture is a significant mediator via children-of-immigrant studies. [More](#)
- I find that many of the papers manage to find statistically significant results. A naive aggregation of the estimated correlation effect sizes suggests that future correlational studies might find effects of around  $\beta \approx 0.28$  (0.13) standard deviations per standard deviation of exposure variation. That is, future studies in similar topics should expect to find that one standard deviation of variation on an event would predict ~28% of a standard deviation in long term outcomes. However it is hard to rule out spurious correlations due to issues such as spatial autocorrelation or outliers. [More](#)
- Some of the papers attempt to study causation via natural experiments. While a couple of such papers arguably succeed in identifying a causal effect, we cannot discard that subsequent robustness checks will cast doubt on the results. A naive aggregation of the estimated correlation effect sizes suggests that future causal studies might find effects of around  $\beta \approx 0.11$  (0.02) standard deviations per standard deviation of exposure variation. That is, future studies in similar topics should expect to find that one standard deviation of difference on an event would cause ~11% of a standard deviation in long term outcomes. [More](#)
- Children-of-immigrant analyses suggest the possibility of parentally mediated long term persistence. The authors of the papers tend to explain this persistence in terms of cultural transmission, relying mostly on historical accounts as evidence. [More](#)
- Whether long term persistence of variation usually stays constant, wanes or increases with time is an open question. Understanding these dynamics of persistence would be critical to understand the long-term impact of cultural interventions today. [More](#)

<b>1. Introduction</b>	<b>4</b>
<b>2. The anatomy of a cultural persistence study</b>	<b>5</b>
2.1 Identifying robust long term correlations	5
2.2 Identifying causal effects	6
2.3 Identifying mediation mechanisms	7
2.4 A gold standard for studying cultural persistence	8
<b>3. Conclusion</b>	<b>8</b>
3.1 Can we rigorously study long-term persistence?	10
3.2 Are the long term correlations studied causal?	13
3.3 Is cultural transmission an important vehicle for persistence?	15
3.4 How does the strength of long term persistence vary through time?	18
3.5 In summary	20
<b>4. Individual reviews</b>	<b>21</b>
<b>The long-term effects of Africa’s slave trades</b>	<b>22</b>
<b>The Slave Trade and the Origins of Mistrust in Africa</b>	<b>26</b>
<b>On the Origins of Gender Roles: Women and the Plough</b>	<b>30</b>
<b>The Church, intensive kinship, and global psychological variation</b>	<b>33</b>
<b>Persecution perpetuated: The medieval origins of anti-semitic violence in nazi Germany</b>	<b>36</b>
<b>Trade, Institutions, and Ethnic Tolerance: Evidence from South Asia</b>	<b>39</b>
<b>Long-term persistence</b>	<b>42</b>
<b>Acknowledgements</b>	<b>44</b>
Bibliography	45
<b>Appendix A: Explanation of the table summary</b>	<b>48</b>
<b>Appendix B: Standard error adjustment for multiple hypothesis testing</b>	<b>51</b>

# 1. Introduction

In the last two dozen years we have seen many papers employing an econometric approach to understanding how events in the long-term past may affect modern cultural outcomes (a phenomenon termed ‘persistence’).

For example, the study of intergenerational persistence includes studying the relation between Africa regions more heavily affected by slavery and their current levels of wealth (Nunn, 2008), the relationship between ancestral usage of plows and modern attitudes towards gender (Alesina et al., 2013) or the relationship between the historical presence of the western church and modern measures of individualism (Schulz et al., 2019).

I review a selection of papers studying these long term correlations, trying to answer three key questions:

- Is there evidence of intergenerational persistence of historical variation?
- Is it possible to identify long term cause-effect relations from historical data?
- Is there evidence of culturally-mediated persistence?

The papers I review are:

- [\*The long term effects of Africa’s slave trades\*](#) (Nunn, 2008)
- [\*The slave trade and the origins of mistrust in Africa\*](#) (Nunn & Wantchekon, 2011)
- [\*On the Origins of Gender Roles: Women and the Plough\*](#) (Alesina et al., 2013)
- [\*The Church, intensive kinship, and global psychological variation\*](#) (Schulz et al., 2019)
- [\*Persecution perpetuated: The Medieval Origins of Anti-Semitic Nazi Violence\*](#) (Voigtländer & Voth, 2012)
- [\*Trade, Institutions, and Ethnic Tolerance: Evidence from South Asia\*](#) (Jha, 2013)
- [\*Long-term persistence\*](#) (Guiso et al., 2016)

## Reading guide

This review is structured as follows:

- [Section 2](#) explains my methodology, outlining a picture of the ideal persistence study and some pitfalls to watch out for.
- [Section 3](#) explains my conclusions, pooling together the results of the individual reviews to answer my research questions.
- [Section 4](#) consists of short reviews for each of the papers. Each review has a standardized table summary<sup>1</sup> containing the results of a replication<sup>2</sup> of an analysis

---

<sup>1</sup> The explanation of the metrics reported in the table summaries can be found in [appendix A](#)

<sup>2</sup> The code to replicate the results is available in <https://github.com/Jsevillamol/persistence>

from the paper<sup>3</sup> and a short conclusion. The replication of additional results can be found in footnotes.

My recommendation for the reader is to skim the first section and read the conclusion. Then the reader might peruse the short reviews of the results that draw their interest.

## 2. The anatomy of a cultural persistence study

Most of the papers I review follow a three-step formula to studying cultural persistence:

1. Identify a long term correlation between a historical measure (the exposure) and a modern one (the outcome)
2. Establish a causal relationship between the exposure and the outcome
3. Establish that the causal effect was mediated via cultural transmission

Let us briefly see how those steps can be approached - as well as how to diagnose some subtle pitfalls<sup>4</sup>.

### 2.1 Identifying robust long term correlations

First of all, we want to **identify a long term correlation**. Identifying statistical correlations in general is a well understood problem for which we have developed multiple statistical tools, the most common one being regression analysis.

Regression analysis allows to easily analyse linear correlations, report the results in a standardised form (as a confidence interval estimating the standardised regression coefficient  $\beta$  or Cohen's  $d$  for example) and control for covariates by including additional regression terms to study the robustness of the results.

One common pitfall we want to watch out for when looking into this study are statistical errors caused by **publication bias**. These manifest primarily in two ways: underpowered studies and multiple hypotheses testing.

**Underpowered studies** introduce a bias that overstates the effect sizes when the results are selected for statistical significance (Button et al., 2013).

To assess *a posteriori* the power of study, I follow (Gelman & Carlin, 2014) and compute the S-value (probability that the result has the wrong sign) and exaggeration ratio (average ratio of the reported and true effect size). This is a highly subjective approach, that requires us to make a judgement call on the expected effect size. For this, I try to refer to previous literature studying similar outcomes at the same level of observation, but the comparison is not perfect.

**Multiple hypothesis testing** (MHT) inflates statistical significance by selecting the most appealing result out of a suite of analyses. To correct the reported p-values for MHT I employ

---

<sup>3</sup> On some cases I report my own analysis, using data from the respective papers

<sup>4</sup> See [appendix A](#) for more information about each of the diagnostic procedures I discuss in this section

the Šidák correction (Šidák, 1967). To adjust the standard error I derive an adjustment formula based on the Šidák correction<sup>5</sup>.

Applying these corrections requires me to make a subjective call on the number of hypotheses being tested. As a conservative lower bound on the number of hypotheses tested I use the number of outcomes studied in a paper. To help the reader make their own judgement on whether this guess is too conservative I also report the critical number of hypotheses that would need to have been tested in order to make the significance of a result fall above the  $p=5\%$  threshold after the correction. Intuitively, this number is a lower bound on the amount of cherry picking that an author would have had to engage in to obtain the reported significance from an insignificant result.

A second pitfall is not controlling properly for **spatial autocorrelation** (SAC). Geographical measures usually cluster together<sup>6</sup>. For example, places nearby have similar altitudes, and furthermore people who live nearby tend to share the same religious beliefs and other cultural traits. (Kelly, 2019) shows how this common spatial distribution of geographical measures induces spurious correlations between them, that we refer to as SAC bias<sup>7</sup>.

To assess the possibility that the reported results are driven by SAC bias we can look at the SAC of the regression residuals using the Moran statistic as suggested by (Kelly, 2019)<sup>8</sup>.

This is not a perfect metric, because the existence of residual SAC does not necessarily mean that the result reported is incorrect. A more comprehensive analysis would attempt to correct for SAC bias with a proper adjustment strategy. One interesting strategy employed by one of the papers I review is to match treatment and control observations based on their geographical location (Voightländer & Voth, 2012). See (Dormann et al., 2007) for a review of some further strategies.

## 2.2 Identifying causal effects

Once we have established a robust correlation between the historical exposure and the long-term outcome we need to **distinguish whether the correlation is indicative of a causal relation** or reflects a common cause<sup>9</sup>.

---

<sup>5</sup> A somewhat handwavy derivation of the adjustment formula can be found in [appendix B](#)

<sup>6</sup> In the words of Waldo Tobler: "everything is related to everything else, but near things are more related than distant things." (Tobler, 2005)

<sup>7</sup> Kelly also shows how the commonly reported Conley standard errors that try to account for this are not sufficient to address SAC bias.

<sup>8</sup> To extend Kelly's analysis to the case with clustered observations we can look at the Moran statistic of the average residuals within each cluster.

<sup>9</sup> Given the strong temporal ordering of the observations, we can rule out reverse causation in these studies. In other domains this would be a concern.

The gold standard for establishing causation is to identify a **natural experiment** - a situation where the value of the historical exposure was (at least partially) assigned in such a way that the assignment itself did not have an effect on the long term outcome.

Formally, we are interested in identifying an instrumental variable - a covariate of the exposure and the outcome such that (1) the instrumental variable is a good predictor of the exposure and (2) the instrumental variable is independent of the long-term outcome when we control for the exposure (the exclusion restriction). Once such an instrument is identified, we can use it to uncover the causal effect of the exposure on the outcome using a two stage least squares analysis (2SLS).

These analyses fail when either of those conditions are violated. We can check whether the instruments chosen are strong enough by computing the F-stat of the first stage. Determining whether the exclusion restriction is violated requires in most cases a subjective call.

Finding relevant natural experiments is not always possible - short of that, the next best option is to ground the hypothetical causal relationship on a theoretical basis and try a battery of regressions under different sets of controls to attempt to falsify the theoretical predictions.

This approach is not ideal because it risks leaving out important confounders (aka omitted variable bias)<sup>10</sup> and furthermore some controls will actually bias the result when included in the regression<sup>11</sup>.

## 2.3 Identifying mediation mechanisms

Once a causal relationship is established, we can focus on trying to **understand if the causal effect was mediated by cultural transmission** or some external factor. For example, in (Nunn & Wantchekon, 2011) the authors investigate whether African slave trade resulted in people who were less trusting or institutions that were less trustworthy.

The gold standard for mediation analysis in this context is **children-of-immigrant studies** ; that is, a study of people living in the same place whose parents came from the places where the historical exposure is measured. This approach implicitly controls all external factors, so that we focus on studying parentally-mediated transmission alone<sup>12 13</sup>.

The main alternative is identifying and controlling for external mediators to see how much they affect the overall result. As before, this strategy is imperfect because we risk leaving important mediators out, but it can help us understand how important external mediators are in the typical case<sup>14</sup>.

---

<sup>10</sup> One partial test for omitted variable bias is selection on observables (Altonji et al., 2005)

<sup>11</sup> through mediator bias, collider bias and M-bias for example (Barret, 2020)

<sup>12</sup> Note that this approach cannot distinguish between purely cultural mechanisms of persistence like parenting and eg genetic mechanisms of persistence

<sup>13</sup> As a bonus, the spatial mixing of immigrants implicitly controls for spatial autocorrelation bias to the extent that the migration is a spatial mixing of the population

<sup>14</sup> For example, if controlling for a set of external mediator reduces the effect to 10% that puts an upper bound to the effect that cultural persistence might exert

## 2.4 A gold standard for studying cultural persistence

So this exercise gives us an idea of what an ideal cultural persistence study looks like:

(1) it identifies a long term correlation with a meaningful effect size and a well-powered sample size, which is robust to additional controls and is not caused by a spurious spatial correlation.

(2) it establishes causality via a natural experiment, or at the very least argues convincingly why a causal relationship should exist based on widely accepted theory

(3) it can replicate its results in a children-of-immigrants study that isolates the cultural transmission effect from other confounders.

As it is often in these cases, this is not a definite recipe, but I found this framework a useful one to understand better the lessons that could be drawn from each paper by comparing how it differed from this made-up ideal.

## 3. Conclusion

For each of the papers I reviewed, I have asked three key questions: 1) How robustly does it establish the correlation it studies? 2) How robustly does it establish a causal relation? 3) How thoroughly does it study the mechanisms that mediate the correlation? Table 1 summarises my findings.

Using the papers as evidence, I then discuss 1) whether we can reliably [identify long term correlations](#), 2) whether we can reliably [identify causal relationships](#) in this context, 3) whether we can assess the [importance of culture for establishing persistent long term effects](#) and 4) whether we should [expect the effect of long term persistence to stay constant](#) or wane as time passes by.

<b>Paper</b>	<b>Theory of causation</b> (Historical exposure ⇒ mediator ⇒ long term outcome)	<b>Correlation analysis</b> Is there a non-spurious long term correlation?	<b>Causality determination</b> Is the correlation causal?	<b>Mediation analysis</b> Can we identify the mechanism of persistence?
<a href="#"><i>The long term effects of Africa's slave trades</i></a> (Nunn, 2008)	Slave trade intensity ⇒ Persistent Ethnic Fractionalization ⇒ GDP per capita	OLS ; severely underpowered, low resilience to MHT adjustment	2SLS (instrument : distance to coast) ; controlling for precolonial prosperity makes the effect disappear	Correlation analysis with hypothesised mediators + theory; correlations are likely to be confounded
<a href="#"><i>The slave trade and the origins of mistrust in Africa</i></a> (Nunn & Wantchekon, 2011)	Slave trade intensity ⇒ Culture ⇒ Trust	OLS ; possible spatial autocorrelation bias	2SLS (instrument : distance to coast); possible spatial autocorrelation bias	OLS studying ancestral exposure <sup>15</sup> , with controls for current historical exposure
<a href="#"><i>On the Origins of Gender Roles: Women and the Plough</i></a> (Alesina et al., 2013)	Ancestral use of plough ⇒ Persistent differentiation of gender roles ⇒ Female labour force participation	OLS ; possible collider bias, not significant after MHT correction	2SLS (instrument : land suitability) ; middling significance, concern that modern measures of the instrument might differ from historical values	Children-of-immigrants OLS ; possible collider bias, not significant after MHT correction
<a href="#"><i>The Church, intensive kinship, and global psychological variation</i></a> (Schulz et al., 2019)	Historical length of exposure to western Church ⇒ Dismantling of kin-based institutions ⇒ WEIRDness	OLS ; statistically and economically significant, no SAC, resilient to MHT	Historical, psychological and anthropological argumentation; no quantitative assessment	Children-of-immigrants OLS
<a href="#"><i>Persecution perpetuated: The Medieval Origins of Anti-Semitic Nazi Violence</i></a> (Voigtländer & Voth, 2012)	Post Black Death antisemitic violence ⇐ Regional persistent cultural antisemitism ⇒ Interwar antisemitic violence	OLS ; only one outcome significant after MHT correction	Theoretical argumentation citing previous work; no quantitative assessment	Interaction analysis ; likely to be underpowered, results not robust
<a href="#"><i>Trade, Institutions, and Ethnic Tolerance: Evidence from South Asia</i></a> (Jha, 2013)	South-asian mediaeval ports ⇒ Interethnic complementarities ⇒ Creation of institutions promoting interethnic cooperation ⇒ lower conflict between muslims and hindus	OLS ; dataset not publicly available so could not analyse in depth	2SLS (instrument: natural harbours) ; results not statistically significant	Interaction analysis ; likely to be underpowered , interpretation not straightforward <sup>16</sup>
<a href="#"><i>Long-term persistence</i></a> (Guiso et al., 2016)	Mediaeval free italian cities ⇒ Self-efficacy culture ⇒ Higher modern civic capital	OLS ; not all results are robust <sup>17</sup> , but one outcome survives MHT correction	2SLS (instrument: presence of a city bishop) ; some issues with the results <sup>18</sup>	OLS study of self efficacy + theory ; I falsify the proposed mechanism through 2SLS

<sup>15</sup> Implicitly this means that the study is studying immigrants and descendent of immigrants, since that is the population where there is variation between the ancestral and geographical historical exposure

<sup>16</sup> This could reflect a lack of understanding on my part

<sup>17</sup> main result based on weighted OLS is not reliable, could not replicate the reported results for the unweighted OLS, small effect sizes on some outcomes, substantial evidence of spatial autocorrelation on some outcomes

<sup>18</sup> Could not replicate some of the results, evidence of spatial autocorrelation bias on some others. There is an outcome that withstands scrutiny though: receiving a gold medal for heroic resistance against the fascist regime



Table 1: Summary of my review of the papers.

### 3.1 Can we rigorously study long-term persistence?

Paper	Identified persistent correlation <sup>19</sup>	Exaggeration ratio	Critical multiple hypothesis number	Residual spatial autocorrelation Moran's Z	Persistence span	Evidence
<a href="#">The long term effects of Africa's slave trades</a> (Nunn, 2008) <sup>20</sup>	-0.49 (0.17) [0.17] +***	3.18	8	-0.64	1400~1900 to 2000	Moderate ; underpowered study
<a href="#">The slave trade and the origins of mistrust in Africa</a> (Nunn & Wantchekon, 2011) <sup>21</sup>	-0.19 (0.03) [0.03] ***	1.01	57.1M	1.76***	1400~1900 to 2005	Moderate ; possible spatial autocorrelation bias
<a href="#">On the Origins of Gender Roles: Women and the Plough</a> (Alesina et al., 2013) <sup>22</sup>	-0.32 (0.15) [0.27]	9.36	2	2.22***	1800~1950 to 2000	Weak ; not robust to MHT
<a href="#">The Church, intensive kinship, and global psychological variation</a> (Schulz et al., 2019) <sup>23</sup>	0.49 (0.07) [0.11] ***	1.19	648M	-0.19	500~1500 to 2014	Strong ; no causal analysis
<a href="#">Persecution perpetuated: The Medieval Origins of Anti-Semitic Nazi Violence</a> (Voightländer & Voth, 2012) <sup>24</sup>	0.11 (0.04) [0.08]**	1.59	7	0.53	1348~1350 to 1920~1945	Moderate ; only one of the studied outcomes is robustly correlated
<a href="#">Trade, Institutions, and Ethnic Tolerance: Evidence from South Asia</a> (Jha, 2013) <sup>25</sup>	-0.17 (0.04) [0.07]***	1.02	572	N/A <sup>26</sup>	700~1700 to 1850~1950	Moderate; dataset not public

<sup>19</sup> Note that these effects are not necessarily causal, and not necessarily culturally mediated.

<sup>20</sup> [https://github.com/Jsevillamol/persistence/blob/master/scripts/1a-slave\\_trade\\_nunn\\_2008.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/1a-slave_trade_nunn_2008.R)

<sup>21</sup> [https://github.com/Jsevillamol/persistence/blob/master/scripts/2a-nunn\\_wantchekon\\_2011-AER.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/2a-nunn_wantchekon_2011-AER.R)

<sup>22</sup> [https://github.com/Jsevillamol/persistence/blob/master/scripts/3a-alesina\\_giuliano\\_nunn\\_2013-country\\_level.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/3a-alesina_giuliano_nunn_2013-country_level.R)

<sup>23</sup> [https://github.com/Jsevillamol/persistence/blob/master/scripts/4a-schulz\\_et\\_al\\_2019-country-level.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/4a-schulz_et_al_2019-country-level.R)

<sup>24</sup> [https://github.com/Jsevillamol/persistence/blob/master/scripts/5-voightlander\\_and\\_voth\\_2012.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/5-voightlander_and_voth_2012.R)

<sup>25</sup> [https://github.com/Jsevillamol/persistence/blob/master/scripts/6a-jha\\_2013.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/6a-jha_2013.R)

<sup>26</sup> The dataset for this paper is not public, so I could not conduct a residual spatial autocorrelation analysis.

<a href="#">Long-term persistence</a> (Guiso et al., 2016) <sup>27</sup>	0.37 (0.01) [0.02] <sup>***</sup>	1.01	$\infty$ <sup>28</sup>	0.30*	1176 to 1943~1945	Moderate ; main outcomes studied are not robustly correlated
---	--------------------------------------	------	------------------------	-------	----------------------	--

Table 2: Summary of evidence for long term persistence. I report in parenthesis () the standard error before MHT correction and in brackets [] the standard error after MHT correction. The asterisk indicate the confidence level after MHT correction: \*=p<10%, \*\*=p<5%, \*\*\*=p<1%. Daggers (†) indicate where I made a significant modification to the analysis conducted in the corresponding paper.

In table 2 we compare some reported effect sizes of the identified correlation. This is not an apples-to-apples comparison: each paper is studying a different pair of exposure and outcome over different time spans, and I am comparing the results of different types of analysis.

Also note that to determine whether the relation was significant we mostly do not care whether their preferred causal history is right. As long as the correlation is not the product of statistical noise or spatial autocorrelation bias it is evidence of *some* persistence mechanism - even if it is not the one the authors had in mind.

All in all, I find that all papers I review provide some evidence in favour of long term persistence<sup>29</sup>. This does not mean that all results in each paper do meet my bar - I have explicitly cherry-picked the results that better withstood my criticism.

Of all the papers, I find that [The Church, intensive kinship, and global psychological variation](#) (Schulz et al., 2019) provides the cleanest case for long-term persistence. It has withstood my diagnosis of statistical power, resilience to multiple hypothesis testing and spatial correlation bias. Its biggest weakness is that its explanation of the causal relation studied mostly relies on previous literature.

[The slave trade and the origins of mistrust in Africa](#) (Nunn & Wantchekon, 2011) is also compelling in its identification of a statistically significant and economically meaningful effect - it identifies an effect robust to alternative measures and sets of controls, and backs it up with a strong causal and mediation analysis. However, there is some room for discussion on whether spatial autocorrelation or multiple hypothesis testing might be inflating the reported results. Indeed, (Kelly, 2020) notes that the results are contingent on the inclusion of respondents from Benin, Ghana and Nigeria. Since respondents from these areas constitute a large part of the sample, I do still trust the result, but it makes me more hesitant. It does not help either that we cannot rule out spatial autocorrelation bias.

<sup>27</sup> [https://github.com/Jsevillamol/persistence/blob/master/scripts/7a-guiso\\_et\\_al\\_2016.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/7a-guiso_et_al_2016.R)

<sup>28</sup> It's not literally infinite but the result is so high it triggers an overflow.

<sup>29</sup> (Jha, 2013) is an edge case, since while their reported results look promising their dataset is not publicly available, and thus I could not analyze it as in depth as I would have liked.

[The long term effects of Africa's slave trades](#) (Nunn, 2008) also identifies a very significant effect - but its analysis is arguably underpowered and it is quite sensible to multiple hypothesis testing. This means that the effect might be smaller than the author reports, and that we need to take on faith that the author did not engage in any cherry picking, intentional or not.

[On the Origins of Gender Roles: Women and the Plough](#) (Alesina et al., 2013), [Persecution perpetuated: The Medieval Origins of Anti-Semitic Nazi Violence](#) (Voigtländer & Voth, 2012) and [Long-term persistence](#) (Guiso et al., 2016) are a mixed bag. One of the outcomes studied on each of the papers seems to exhibit long term persistence<sup>30</sup>, but closely related outcomes and analysis do not hold up. After correcting for multiple hypothesis testing, only (Guiso et al., 2016) analysis of the gold medal outcome remains strongly statistically significant.

In brief, I believe that these results do make a compelling and rigorous case for the existence of long term correlations. Some of the minimum significance levels that robustly stand many alternate analyses are well below the range where we should expect random noise to be confused by actual patterns.

Having established that long term persistence is likely to exist, can we make some estimations on how strong it is?

Not with confidence. The number of papers is small and their confidence intervals are wide. Of some of the papers we suspect (on priors) that are overestimating the effect size, as estimated by the exaggeration ratio.

A very naive pooling<sup>31</sup> of the absolute estimated effects<sup>32</sup> of the selected analysis suggests a pooled correlation effect size of about  $\beta \approx 0.28$  (0.13)<sup>33</sup>.

Taking this blindly, this suggests that a reasonable prior is to expect that an observed historical variation of one standard deviation on those cases where we can reliably detect<sup>34</sup> long term persistence is associated with  $\sim 1/4$  of a standard deviation in modern outcomes on expectation.

---

<sup>30</sup> Those would respectively be historical agricultural practices and female labor force participation, medieval and interwar antisemitic violence in Germany and medieval independence status of Italian cities and condecorations for resisting the fascist regime in 1940s.

<sup>31</sup> Details on meta-analytical method:

- Inverse variance method
- DerSimonian-Laird estimator for  $\tau^2$
- Jackson method for confidence interval of  $\tau^2$  and  $\tau$
- Hartung-Knapp adjustment for random effects model (Borenstein, 2009; M.Sc.<sup>1</sup> et al., n.d.)

<sup>32</sup> after MHT correction.

<sup>33</sup> <https://github.com/Jsevillamol/persistence/blob/master/scripts/meta.R>

<sup>34</sup> The wording is important - there might be smaller persistent effects that we cannot reliably detect due to statistical power issues.

This estimate does not take into account that a big part of the association is spurious and not causal. We will address this in the next section.

### 3.2 Are the long term correlations studied causal?

Paper	Identified persistent causal effect <sup>35</sup>	Persistence span	Evidence
<a href="#">The long term effects of Africa's slave trades</a> (Nunn, 2008) <sup>36</sup>	-0.36 (0.34) [0.34] <sup>†</sup>	1400~1900 to 2000	Weak ; result underpowered
<a href="#">The slave trade and the origins of mistrust in Africa</a> (Nunn & Wantchekon, 2011) <sup>37</sup>	-0.10 (0.03) [0.03] <sup>***</sup>	1400~1900 to 2005	Moderate ; possible spatial autocorrelation bias
<a href="#">On the Origins of Gender Roles: Women and the Plough</a> (Alesina et al., 2013) <sup>38</sup>	-0.78 (0.37) [0.66]	1800~1950 to 2000	Moderate ; not significant after MHT adjustment, weak concern about validity instrument
<a href="#">The Church, intensive kinship, and global psychological variation</a> (Schulz et al., 2019)	N/A	500~1500 to 2014	Weak ; argument is mainly theoretical
<a href="#">Persecution perpetuated: The Medieval Origins of Anti-Semitic Nazi Violence</a> (Voigtländer & Voth, 2012)	N/A	1348~1350 to 1920~1945	Weak ; argumentation is mainly theoretical
<a href="#">Trade, Institutions, and Ethnic Tolerance: Evidence from South Asia</a> (Jha, 2013) <sup>39</sup>	-0.16 (0.34) [0.56]	700~1700 to 1850~1950	Weak ; IV analysis not significant
<a href="#">Long-term persistence</a> (Guiso et al., 2016) <sup>40</sup>	0.11 (0.02) [0.04] <sup>†***</sup>	1176 to 1943~1945	Moderate ; original result

Table 3: Evidence of causality. The effect sizes reported correspond to the results of natural experiments. I report in parenthesis () the standard error before MHT correction and in brackets [] the standard error after MHT correction. The asterisk indicate the confidence level after MHT correction: \*=p<10%, \*\*=p<5%, \*\*\*=p<1%. Daggers (†) indicate where I made a significant modification to the analysis conducted in the corresponding paper.

[The slave trade and the origins of mistrust in Africa](#) (Nunn & Wantchekon, 2011) and [Long-term persistence](#) (Guiso et al., 2016) both provide some evidence of an actually causal effect in the correlations they study through their study of natural experiments.

<sup>35</sup> Note that these effects are not necessarily culturally mediated

<sup>36</sup> [https://github.com/Jsevillamol/persistence/blob/master/scripts/1b-slave\\_trade\\_nunn\\_2008-IV.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/1b-slave_trade_nunn_2008-IV.R)

<sup>37</sup> [https://github.com/Jsevillamol/persistence/blob/master/scripts/2b-nunn\\_wantchekon\\_2011-IV.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/2b-nunn_wantchekon_2011-IV.R)

<sup>38</sup> [https://github.com/Jsevillamol/persistence/blob/master/scripts/3b-alesina\\_giuliano\\_nunn\\_2013-country\\_level\\_IV.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/3b-alesina_giuliano_nunn_2013-country_level_IV.R)

<sup>39</sup> [https://github.com/Jsevillamol/persistence/blob/master/scripts/6b-jha\\_2013-IV.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/6b-jha_2013-IV.R)

<sup>40</sup> [https://github.com/Jsevillamol/persistence/blob/master/scripts/7b-guiso\\_et\\_al\\_2016-IV.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/7b-guiso_et_al_2016-IV.R)

Of these, [The slave trade and the origins of mistrust in Africa](#) (Nunn & Wantchekon, 2011) is robust to all the checks I performed, including an additional IV analysis that includes terms for both ancestral and geographical exposure to slavery. Like in (Kelly, 2020), I find that the result is no longer significant if we remove respondents from Benin, Ghana and Nigeria. But given these respondents are a large part of the sample I think it is still valid. However, the high degree of spatial autocorrelation is a concern.

In [Long-term persistence](#) (Guiso et al., 2016) the authors manage to identify an arguably suitable instrument -the existence of a bishop that facilitated coordinating the independization of the city- but I don't find the outcomes studied by the authors to have a robust correlation with the exposure. However, the authors also collect data about whether the cities received a gold medal for resisting the fascist regime in 1943-1945, and while they do not subject it to the IV analysis when I do so I find a statistically and economically significant effect that withstands MHT correction. Given that this is an original result and that the rest of the proxies didn't fare well against MHT correction I am cautiously optimistic but not convinced about the validity of this result.

[The long term effects of Africa's slave trades](#) (Nunn, 2008), [On the Origins of Gender Roles: Women and the Plough](#) (Alesina et al., 2013) and [Trade, Institutions, and Ethnic Tolerance: Evidence from South Asia](#) (Jha, 2013) also attempt to quantitatively study causality through a natural experiment, however I find their results inconclusive.

While the main IV result reported [The long term effects of Africa's slave trades](#) (Nunn, 2008) in the absence of controls is significant, after controlling for certain geographical controls the relation weakens and ceases to be statistically significant.

The natural experiment analysis in [On the Origins of Gender Roles: Women and the Plough](#) (Alesina et al., 2013) does not withstand MHT correction. There is also some room for discussing whether the instrument used is suitable. The modern measure of the instrument -suitability of land to different crops- might differ substantially from the historical case<sup>41</sup>.

In [Trade, Institutions, and Ethnic Tolerance: Evidence from South Asia](#) (Jha, 2013) the results reported are outright not significant.

All in all, I find that some of the papers provide some quantitative evidence of causal long term persistence.

Despite this, I also need to admit that the quantitative evidence is not robust - upon closer inspection, many of the results that attempted to demonstrate causation are weaker than initially expected.

Better tools for studying causation quantitatively will be needed before a conclusion is reached.

Putting that aside, can we make a guess on how strong are the causal effects we might expect to find in future persistence studies?

---

<sup>41</sup> I have substantial uncertainty here, since I do not have a good grasp of how continued land use changes agricultural suitability

If we naively pool the causal effect sizes<sup>42</sup> in table 3 we find an aggregate effect of  $\beta \approx 0.11$  (0.02)<sup>43</sup>.

We can interpret this as a naive estimate of the long term causal effects we might be able to identify in similar contexts.

Having identified persistent causal effects is different from understanding the mechanisms through which these effects persist. We address this in the next section.

### 3.3 Is cultural transmission an important vehicle for persistence?

One question the papers I review try to address is identifying the most important mechanisms that mediate these long-term correlations.

The main mechanism studied in the papers and the one I care most about for the purposes of this review is culture.

What exactly is cultural transmission is something hard to pinpoint, but I use it to refer to the cluster of mechanisms like parenting and word-of-mouth, rather than things like the institution of organisations or genetics.

As an inexact operationalization, I refer to cultural mechanisms as those attitudes and beliefs that a person would carry with them when they emigrate to another country, and would be handed down from parents to children.

Negative examples of this narrow definition of culture include for example governments and constitutions.

In practice, most papers I reviewed resort to qualitative accounts to justify their favoured explanations of persistence. Some of them complement those results with quantitative studies of the persistence effect of parentage - however we should bear in mind this strategy cannot distinguish between cultural and non cultural mechanisms of parentage<sup>44</sup>.

Paper	Favoured explanation of persistence	Identified parentage effect size <sup>45</sup>	Evidence
<a href="#"><i>The long term effects of Africa's slave trades</i></a> (Nunn, 2008)	Cultural	N/A	Weak ; unconvincing mediation analysis
<a href="#"><i>The slave trade and the origins of mistrust in</i></a>	Cultural	-0.19 (0.03) [0.03] ***	Moderate ; doubts about the effects of outliers

<sup>42</sup> after multiple hypothesis testing correction

<sup>43</sup> <https://github.com/Jsevillamol/persistence/blob/master/scripts/meta.R>

<sup>44</sup> Eg parental gene transmission

<sup>45</sup> Not necessarily causal, and not necessarily cultural

<a href="#">Africa</a> (Nunn & Wantchekon, 2011) <sup>46</sup>			
<a href="#">On the Origins of Gender Roles: Women and the Plough</a> (Alesina et al., 2013) <sup>47</sup>	Cultural	-0.01 (0.01) [0.03]	Weak ; insignificant after correcting for collider bias
<a href="#">The Church, intensive kinship, and global psychological variation</a> (Schulz et al., 2019) <sup>48</sup>	Cultural	0.12 (0.01) [0.03] ***	Strong ; robust children-of-immigrant analysis, though no quantitative causal analysis
<a href="#">Persecution perpetuated: The Medieval Origins of Anti-Semitic Nazi Violence</a> (Voigtländer & Voth, 2012)	Cultural	N/A	Weak ; unconvincing interaction analysis
<a href="#">Trade, Institutions, and Ethnic Tolerance: Evidence from South Asia</a> (Jha, 2013)	Institutional	N/A	Weak ; unconvincing interaction analysis
<a href="#">Long-term persistence</a> (Guiso et al., 2016)	Cultural	N/A	Weak ; I falsify their cultural mediation hypothesis

Table 4: Evidence for mediation mechanism of long-term persistence. The effect sizes reported correspond to the results of children-of-immigrant analysis. I report in parenthesis () the standard error before MHT correction and in brackets [] the standard error after MHT correction. The asterisks indicate the confidence level after MHT correction: \*=p<10%, \*\*=p<5%, \*\*\*=p<1%.

Of the seven papers I review, only two meet my (admittedly high) bar for relevant evidence on cultural persistence.

[The slave trade and the origins of mistrust in Africa](#) (Nunn & Wantchekon, 2011) performs a study correlating modern levels of trust of individuals (as measured by multiple answers in a survey) with the intensity of slave trade in the area inhabited by their ancestors, while controlling among other things for the historical intensity of slave trade in the geographical area where each respondent lives.

<sup>46</sup> [https://github.com/Jsevillamol/persistence/blob/master/scripts/2a-nunn\\_wantchekon\\_2011-AER.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/2a-nunn_wantchekon_2011-AER.R)

<sup>47</sup> [https://github.com/Jsevillamol/persistence/blob/master/scripts/3d-alesina\\_giuliano\\_nunn\\_2013-USA\\_children\\_of\\_immigrants.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/3d-alesina_giuliano_nunn_2013-USA_children_of_immigrants.R)

<sup>48</sup> [https://github.com/Jsevillamol/persistence/blob/master/scripts/4c-schulz\\_et\\_al\\_2019-children\\_of\\_immigrant\\_europe.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/4c-schulz_et_al_2019-children_of_immigrant_europe.R)

This strategy, in my opinion, convincingly isolates the cultural component of this surprising long-term correlation. And the remaining correlation is indeed significant, economically and statistically speaking.

Furthermore, the paper provides compelling causal evidence for the phenomena.

[\*The Church, intensive kinship, and global psychological variation\*](#) (Schulz et al., 2019) cannot manage to find a compelling natural experiment to base their analysis on, but regardless finds a statistically and economically significant long term correlation between length of exposure to the Western church in the country of origin of the parents of European children of immigrants and western attitudes (eg individualism).

This puts this paper in a complex situation - they appeal to a compelling causal narrative, backed up by previous research in psychology, anthropology and history, where variation in western attitudes is mediated by the dismantling of kin based institutions orchestrated by the western church. But I am not in a good position to assess the quality of this previous research. Just going by the results in the paper, we cannot tell whether this explanation is correct - but their analysis indicates that *something* surviving cultural mediation is relating the exposure to the western church to western attitudes.

While [\*On the Origins of Gender Roles: Women and the Plough\*](#) (Alesina et al., 2013) does not provide compelling evidence of cultural persistence, it is worth noting that it arguably manages to identify a causal effect of agricultural practices on female labour participation - and the fact that this effect does not robustly register in children of immigrants suggests that there exists some external factors mediating the correlation. Identifying a case where the long term effect is (arguably) primarily mediated by external factors does not mean that parental mediation does not exist - but it should update us slightly down on how likely are the other papers to have identified primarily culturally mediated correlations<sup>49</sup>.

In brief, I believe that these papers provide some evidence on the existence of cultural differences, originally shaped by happenstance but perpetuated from parent to children through generations. It is however very hard to assess whether the mechanism of transmission was cultural or not, and these papers do not provide quantitative evidence in that respect.

The idea that a non trivial part of our culture may be derived from events that happened centuries of years ago may be shocking to some - and indeed the results studied by (Nunn & Wantchekon, 2011) and (Schulz et al., 2019) are quite striking.

But this feels less impossible when we focus on more familiar examples, those that we take for granted - for example, it is not a stretch to argue that religious practice a thousand years ago still shapes our beliefs and traditions.

---

<sup>49</sup> It is worth stressing again that we are investigating some very particular aspects of culture, those which immigrants carry with them when moving their home country and is transmitted through parentage



### 3.4 How does the strength of long term persistence vary through time?

Another question we are interested in is how the degree of persistence of the correlations studied in the papers changes through time.

We saw previously that in those cases where long term persistence is identified a very naive prior for the (non causal) effect size is about  $\beta \approx 0.28$  (0.13).

Should we expect this persistent variation to stay constant through time? Or does it fade away the further we go?

The papers I reviewed do not study this question in depth, so I do not have a lot of evidence to base claims on.

If we look at the identified persistent correlations and the span of persistence in [Table 2](#) there is no obvious correlation between the length of the persistence window and the identified effect. But 7 data points is not a lot to go on, and the studies are too heterogeneous to draw significant conclusions regardless.

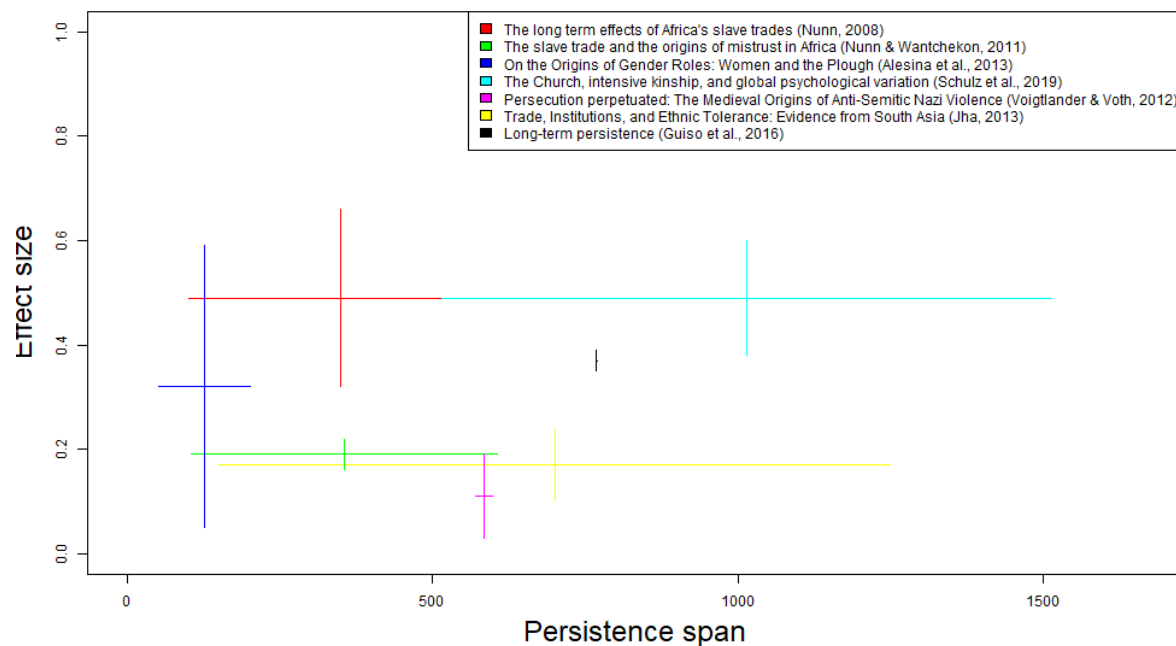


Figure 1: Persistence span vs correlation effect size

A related question is whether longer exposures might lead to stronger effect sizes. Again, plotting the information we have about these two does not indicate any obvious relation between them.

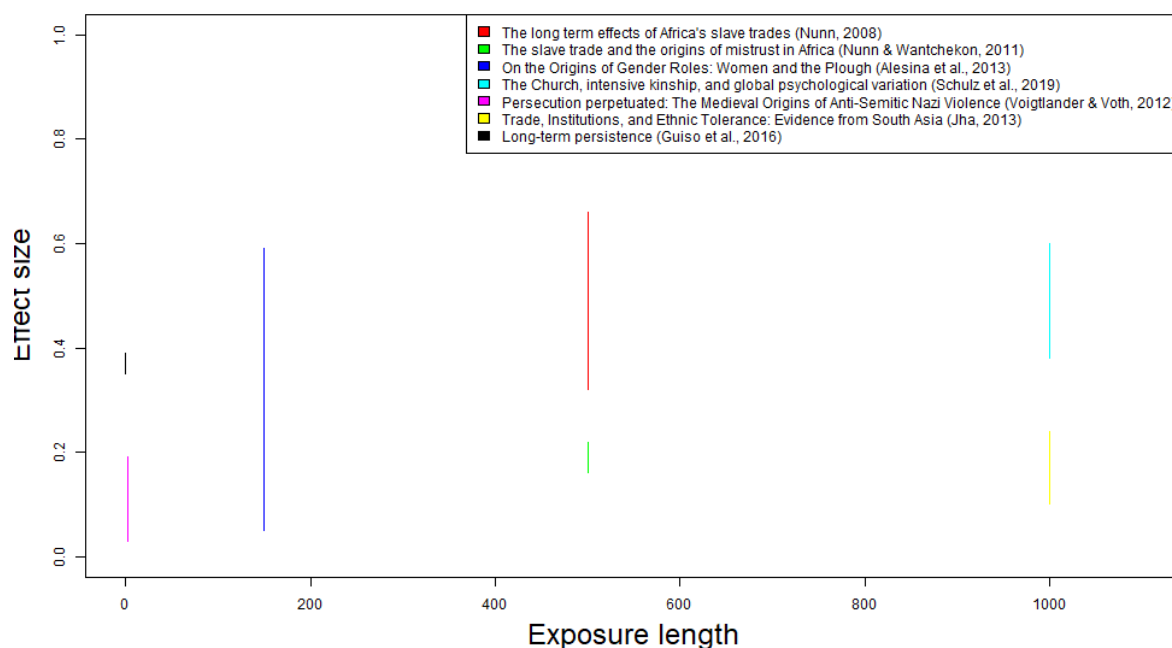


Figure 2: Duration of exposure versus reported correlation effect size.

A better approach might be to look into papers that study outcomes measured at different times.

[\*Persecution perpetuated: The Medieval Origins of Anti-Semitic Nazi Violence\*](#) (Voigtländer & Voth, 2012) relates medieval antisemitic violence following the Black Death in ~1350 to 1920s antisemitism. But they also study other indicators of antisemitism through history, including the Hep Hep riots in the 1800s, and find only a very weak correlation. Taken at face value, this is a strange pattern of persistence - albeit they use this to explain something else, one hypothesis put forward by the author is that the underlying persistent culture only surfaces when the conditions are right. We cannot rule this out, but I am unpersuaded by this explanation.

[\*Long-term persistence\*](#) (Guiso et al., 2016) studies outcomes in the 1943~1945 period and in the 2000s - and my subsequent analysis reveals that only the former are significant. This is evidence of fading, but rather weak - a more plausible explanation might be that the outcomes studied in the 2000s (proxies of civic capital) are significantly different from the outcome studied in 1943~1945 (resistance against the fascist regime).

Summing up, with the evidence at hand we cannot settle the question of whether historical influence stays constant or fades away and how quickly.

This is a question of paramount importance from a longtermist perspective, as its answer might dominate the cost-effectiveness of interventions aiming to produce persistent cultural change.

### 3.5 In summary

The papers I've reviewed provide some evidence that long-term persistence exists and that culture is an important persistent factor.

However, many of the results were far less significant after performing some robustness checks. This, together with the recent work by (Kelly, 2020) suggests that our current tools for quantitatively studying causation are not adequate for the task. Further work is needed before a conclusion can be reached.

On top of that, it is far from clear what are realistic bounds for the effect sizes of long-term cultural persistence, and how the effect size might vary through time. Future research may be interested in looking into these questions more in depth.

## 4. Individual reviews

This section includes the short reviews of the individual papers:

- [\*The long term effects of Africa's slave trades\*](#) (Nunn, 2008)
- [\*The slave trade and the origins of mistrust in Africa\*](#) (Nunn & Wantchekon, 2011)
- [\*On the Origins of Gender Roles: Women and the Plough\*](#) (Alesina et al., 2013)
- [\*The Church, intensive kinship, and global psychological variation\*](#) (Schulz et al., 2019)
- [\*Persecution perpetuated: The Medieval Origins of Anti-Semitic Nazi Violence\*](#) (Voigtländer & Voth, 2012)
- [\*Trade, Institutions, and Ethnic Tolerance: Evidence from South Asia\*](#) (Jha, 2013)
- [\*Long-term persistence\*](#) (Guiso et al., 2016)

For each paper I have included the details of a flagship analysis in a table summary, either replicating or extending the results in the paper. An explanation of the information contained in the tables can be found in [appendix A](#). Additional replications and checks can be found in the footnotes. The code I used to replicate the results is available in <https://github.com/Jsevillamol/persistence>.

# The long-term effects of Africa's slave trades

by Nathan Nunn

## Table summary

<b>Result of replication</b>	Not significant causal correlation between log slave exports normalized by area and log real GDP per capita	
<b>Statistical method of replication</b>	Two stage least squares regression Instrument: distance to the coast	
<b>Standardised <math>\beta \approx -0.36</math> (0.34) [0.34]</b>	<b>Adjusted p-value <math>\approx 30\%</math></b>	<b>n=52 [10 outliers]</b>
<b>Expected <math>\beta \approx -0.13</math><sup>50</sup></b>	<b># hypothesis = 1</b>	<b>Critical number of hypothesis = N/A</b>
<b>Power <math>\approx</math> N/A</b>	<b>Type S error rate <math>\approx</math> N/A</b>	<b>Exaggeration ratio <math>\approx</math> N/A</b>
<b>Moran's Z <math>\approx -0.64</math></b>	<b>Moran's p <math>\approx 52\%</math></b>	<b>Persistence span = 1400~1900 to 2000<sup>51</sup></b>
<b>First stage F-stat = 3.84</b>		
<b>Short commentary</b>	<p>Based on our prior expectations for the likely effect size, I find that Nunn correlation analysis is severely underpowered, and the design is likely to overestimate the effect size by an average factor of 3.</p> <p>Furthermore, I find that the IV analysis fails to control for precolonial prosperity, and once we do so the result stops being statistically significant.</p> <p>Lastly, I think that the causal mechanisms put forward in the paper (ethnic fragmentation and state underdevelopment) are likely to play a small part relative to the effect mediated by the direct decrease in population.</p>	
<b>Reproduction details</b>	<p>We reproduce the analysis in table IV, column (3), with an additional control for pre colonial prosperity.</p> <p><a href="https://github.com/Jsevillamol/persistence/blob/master/scripts/1b-slave_trade_nunn_2008-IV.R">https://github.com/Jsevillamol/persistence/blob/master/scripts/1b-slave_trade_nunn_2008-IV.R</a></p>	

<sup>50</sup> Based on the geometric mean of 11 correlations studied by (Ilter, 2017)

<sup>51</sup> "Figure III shows the relationship between the natural log of the number of slaves exported in all four slave trades between 1400 and 1900 normalized by land area and the natural log of per capita GDP in 2000." (Nunn, 2008)

## Review conclusion

Despite the limited data, Nunn's analysis shows a robust negative correlation between slave trade intensity and modern economic prosperity of African countries. Concretely, Nunn:

- 1) applies several subsets of controls for geography, legal origin natural resources and region, uses different proxies for slave trade intensity, and repeats the analysis excluding outliers in several ways and finds that none of this alternate analysis significantly changes the bottom line.
- 2) studies how the bias in his data that undercounts slave trade in landlock countries affects that data, and finds that the undercounting understates the effect.

We also analyse and find no evidence of spatial autocorrelation among the residuals of the analysis, suggesting that the result is not the consequence of a spurious spatial correlation.

However, we find that this study is likely to be underpowered. (Ilter, 2017) provides a superficial analysis of the relation between 11 factors<sup>52</sup> and GDP per capita. If we take both the significant and non-significant estimates, the geometric average of the standardised effect size of 0.13 $\sigma$ . Using this figure as our expected effect size, we find that Nunn's study is expected to overestimate the effect size by an average factor of 3.<sup>53</sup>

Furthermore, the strategies Nunn pursues to establish causality are suspect. In particular, he:

- 1) uses sailing distance to trade hubs as a (weak) instrumental variable, and this IV analysis suggests that the relationship is stronger than what a naive regression indicates - consistent with the bias introduced by the undercounting of slaves captured in landlocked countries who did not make it to the coast. However, these instruments are suspect -- distance to major trade hubs is surely correlated with economic prosperity in many other ways than as a driver for slave demand. Nunn explains that the instruments are only correlated with modern GDP capita within Africa, and not without Africa, but does not explain how this analysis was carried out.
- 2) if we perform the IV analysis without additional controls we cannot rule out spatial autocorrelation bias<sup>54</sup>, but additional controls for geography and coloniser fixed effects significantly weaken the statistical significance of the result<sup>55</sup>.

---

<sup>52</sup> These factors are population, land area, transparency score, transparency ranking, GDP, current GDP growth, inflation consumer price index, youth unemployment, population below poverty line, compulsory education years, government

<sup>53</sup> "beta = -0.49 (0.17) [0.17], adjusted p = 7.02e-03" "Exaggeration ratio =3.18"

[https://github.com/Jsevillamol/persistence/blob/master/scripts/1a-slave\\_trade\\_nunn\\_2008.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/1a-slave_trade_nunn_2008.R)

<sup>54</sup> I get "Moran's Z = 3.24" [1] "Moran's p = 1.21e-03" when conducting the 2SLS without any controls

[https://github.com/Jsevillamol/persistence/blob/master/scripts/1b-slave\\_trade\\_nunn\\_2008-IV.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/1b-slave_trade_nunn_2008-IV.R)

<sup>55</sup> Without the precolonial prosperity controls I find "beta = -0.52 (0.12) [0.12], adjusted p = 8.30e-05". With the precolonial prosperity controls I find "beta = -0.53 (0.36) [0.36], adjusted p = 1.47e-01"

- 3) studies whether poorer precolonial economies self selected into slave trade -as a possible confounder-, and finds evidence that actually the *richer* pre colonial regions suffered more slave trade. Nunn furthermore assumes a positive correlation between pre colonial and modern prosperity, which together with the self selection effect would push the correlation between slave trade and modern GDP per capita away from 0 if controlled.

This assumption is reasonable at first glance, but previous work has argued for a negative correlation between pre colonial and post colonial prosperity (Acemoglu et al., 2001).

Indeed, controlling for precolonial prosperity<sup>56</sup> reduces the statistical significance of the instrumental variable analysis<sup>57</sup>. The loss of significance is mainly driven by a decreased estimated effect size, not an increase in the standard error - which is evidence in favour of precolonial prosperity being a confounder biasing the effect away from zero, against Nunn's prediction.

The former considerations are not enough to deny Nunn's result -- but they significantly weaken the evidence.

Putting that aside, Nunn puts forward the hypotheses that the effects of slave trade on modern economic prosperity might be mediated by ethnic fractionalization or state underdevelopment.

He supports these hypotheses by studying the correlation of slave trade intensity with proxies of these (which shows a strong positive correlation with ethnic fractionalization and a weak one with state underdevelopment), and offering a historical account of them. But these analyses are carried in the absence of controls or other causal considerations.

All in all, while probably important on some level, it is unclear whether these hypotheses are sufficient to explain the extreme effect sizes found in the paper.

A parsimonious solution to this is just dismissing the effect size as the result of the low power of the analysis.

Alternatively we may want to look into different mediation mechanisms. One that I am surprised was not considered is how slave trade set back the population level of Africa by at least 25 years<sup>58</sup>, denying it of useful labour.

This is far from a knock-down argument, as standard growth economics posits that developing countries are more impeded by an inefficient distribution of labour rather than a

---

<sup>56</sup> Measured as precolonial population, and production of gold, oil and diamonds per capita

<sup>57</sup> Including the precolonial prosperity controls make the significance fall from "beta = -0.53 (0.36) [0.36]" [1] "p = 1.47e-01" to beta = beta = -0.36 (0.34) [0.34], adjusted p = 3.01e-01

[https://github.com/Jsevillamol/persistence/blob/master/scripts/1b-slave\\_trade\\_nunn\\_2008-IV.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/1b-slave_trade_nunn_2008-IV.R)

<sup>58</sup> I derive this figure from Nunn quoting "The total effect of the slave trades, according to calculations by Patrick Manning (1990, p. 171), was that by 1850 Africa's population was only half of what it would have been had the slave trades not taken place." and Africa's top registered historical annual population growth rate of 2.83% (Roser et al., 2013).

shortage of labour *per se*, but the population decrease is so extreme that this hypothesis merits attention - as long as we concede that returns to labour on GDP per capita in African countries are positive, even if low.

One plausible story that would explain why returns to labour might be positive is that having extra labour produces further innovation, and local adaptation of innovation, which are non-rival goods that boost the productivity of the country (Muthukrishna & Henrich, 2016).

In brief, Nunn identifies a correlation between historical slavery and modern prosperity, through an underpowered analysis that likely overstates the effect.

Despite a solid attempt, I do not believe he managed in this paper to convincingly prove that the correlation is causal, nor to identify the mechanisms that mediate the correlation.



# The Slave Trade and the Origins of Mistrust in Africa

by Nathan Nunn and Leonard Wantchekon

## Table summary

<b>Main result of replication</b>	Negative correlation between ethnic slave trade incidence and five indicators of trust <sup>59</sup> , mainly mediated by a factor transmitted by parentage	
<b>Statistical method of replication</b>	IV analysis, regressing on ancestral values of the exposure while controlling for geographical historical values of the exposure Instrument: distance to the coast	
<b>Standardized <math>\beta \approx -0.10</math> (0.03) [0.03]</b>	<b>Adjusted p-value <math>\approx 0.004</math></b>	<b>n =</b> 251 ethnicities 1143 districts 15972 individuals
<b>Expected <math>\beta \approx -0.130^{60}</math></b>	<b># hypothesis = 1<sup>61</sup></b>	<b>Critical hypotheses number = 14</b>
<b>Power <math>\approx 0.97</math></b>	<b>Type S error rate <math>\approx 0\%</math></b>	<b>Exaggeration ratio <math>\approx 1.02</math></b>
<b>Moran's Z <math>\approx 1.76</math></b>	<b>Moran's p <math>\approx 0.078</math></b>	<b>Persistence span = 1400~1900 to 2005<sup>62</sup></b>
<b>Short commentary</b>	<p>Of the many analyses in the paper, the results rest mostly on two: 1) the control for slave trade incidence on region and ethnic ascendants and 2) the analysis based on the instrument of coastal distance.</p> <p>Together they make a case for a causal link between slave trade and trust mediated by a factor transmitted by parentage.</p> <p>It is unclear whether the mechanism is primarily driven by a value shift, or a selection effect.</p>	

<sup>59</sup> Those indicators are derived from the same survey asking questions about trust of relatives, neighbours, local council, intragroup trust and intergroup trust

<sup>60</sup> From (Guiso et al., 2003) we find a geometrically average effect size of the correlation between several variables and trust of  $d \approx 0.26$ . Assuming the treatment to be distributed as a fair coin, this gives an expected standard effect of  $\beta \approx 0.13$ .

<sup>61</sup> While multiple outcomes are studied in the paper, all are closely related (they are self reported trust measures from the same survey) so I do not think a multiple hypothesis adjustment is needed

<sup>62</sup> "The country-level estimates cover Africa's four slave trades (the transatlantic, Indian Ocean, Red Sea, and trans-Saha-ran) between 1400 and 1900. [...] The individual-level data are from the 2005 Afrobarometer surveys." (Nunn & Wantchekon, 2011)

	We also could not rule out spatial autocorrelation bias and multiple hypothesis bias.
<b>Reproduction details</b>	<p>I extend the results in the paper with a 2SLS analysis with outcome trust of neighbours and regression terms for ethnic-level ancestry slave trade intensity and country-level, plus the individual, district, country and ethnic controls specified in the text. The instrument I use is the historic distance of the respondent's ethnic group to the coast.</p> <p>For the standard errors we use clustered standard errors using the <code>lfe::felm</code> function, clustering at the ethnicity and district level.</p> <p>The Moran statistics are computed over the ethnicity-level averages.</p> <p><a href="https://github.com/Jsevillamol/persistence/blob/master/scripts/2b-nunn_wantchekon_2011-IV.R">https://github.com/Jsevillamol/persistence/blob/master/scripts/2b-nunn_wantchekon_2011-IV.R</a></p>

## Review conclusion

Through this paper the authors study the relation between historical slave trade intensity of different African ethnic regions and modern trust of their descendants.

There is much to like about this paper. However there are two analyses that form, in my opinion, the crux of the paper and on whose validity rests the conclusion:

1. The instrumental variables analysis in table 5, with coastal distance as an instrument
2. The OLS regression controlling for both ethnic level slave trade intensity and current-region-of-residence slave trade intensity in table 10

The former establishes a negative causal link between slave trade and modern trust, while the second establishes parentage as the main mediator for the causal effect.

One might be surprised that I put so much trust on the instrument analysis, and in fact a bit of scepticism is warranted - distance to the coast is a significant economic factor with many repercussions, so it calls into question whether the exclusion restriction is being met. But Nunn and Wantchekon show in tables 7 and 8 that only in Africa is coastal distance correlated with trust - and I find this hard to ignore.

The OLS regression distinguishing effects on the region versus effects on families only applies to those who have moved from their ancestral home - but those constitute 43% of the sample. It shows that internal factors are about 3 times as important as external ones.

This analysis does not distinguish the exact mechanism of transmission - of the hypotheses ventured by Nunn and Wantchekon at the beginning of the paper, the ones compatible with this data are either a cultural shift in values as the population learned to distrust possible slavers or a selection effect as the more trusting people were enslaved.

Beyond these two analyses, Nunn and Wantchekon offer a plethora of alternate analyses that add further evidence to their case, but I do not find them as load-bearing.

In table 1 they present an analysis showing how different measures of slave trade intensity and different adjustments of the standard error (including one adjustment for spatial dependency) do not significantly affect their results.

The analysis at tables 2 and 3 are subsumed by the analysis table 10, as variants with less controls and similar results.

In table 4 they use selection of observables to discredit the possibility of weak confounders affecting their analysis. But says nothing about the possibility of missing strong confounders.

Table 6 is the IV analysis from table 5 with additional controls - does not change things significantly.

Table 9 contains two analyses, where the authors attempt to distinguish whether slave trade made Africans less trusting or their environment less trustworthy by controlling for the trustworthiness of the respondents' city councils and the trustworthiness of their other peers. The results are consistent with a mainly internally mediated shift in trust, but I find the proxies chosen weak and thus the results inconclusive.

If I try to come up with the strongest reasons to be sceptical of the result, what I would put in the top is:

- 1) the punctual use of spatial-correlation-adjusted standard errors is not enough to discard the possibility of spatial autocorrelation bias<sup>63</sup> - and indeed we find that the residuals are spatially correlated at the 90% confidence level according to the Moran test. The case for spatial autocorrelation bias is stronger after we consider how (Kelly, 2020) notes that the results of this paper are contingent on the inclusion of respondents from Benin, Ghana and Nigeria - this decreases my confidence in the paper. But these respondent constitute a large part of the sample<sup>64</sup>, so Kelly's critique might be too overzealous.
- 2) the absence of correlation between trust and coastal distance everywhere but in Africa is surprising enough to warrant a second look.
- 3) the study seems underpowered - from (Guiso et al., 2003) we find a geometrically average effect size of the correlation between several variables and trust of  $d \approx 0.26$ . Using this figure as the expected effect size gives the study an exaggeration ratio of  $\sim 3$ . So the effect size is likely overstated.

---

<sup>63</sup> Kelly 2019 argues that the sensitivity of the correction is insufficient: "The fact that adjusted standard errors are largely unchanged is a consequence of setting an extremely low cutoff radius beyond which observations are given a weight of zero. For example Nunn and Wantchekon (2011) set a radius of three degrees for Africa, which spans about 70 degrees each way"

<sup>64</sup> There are 3948 respondents from Benin, Ghana and Nigeria, which constitutes 24.7% of the sample. They come from 365 districts, which are 31.9% of all the districts.

One further qualm we could pose is that while the IV analysis establishes a causal relation, and the ancestral OLS regression controlling for both ethnic level slave trade intensity and current-region-of-residence slave trade intensity establishes a cultural mediation, the aspects that are culturally mediated might not be causal.

To test this, I extend Nunn and Wantchekon's analysis, performing an IV analysis while controlling for ancestral and geographical historical slave intensity. The results remain significant<sup>65</sup>. Same as with Kelly, after removing the respondents from Benin, Ghana and Nigeria the results are no longer significant<sup>66</sup>.

All in all, I am inclined to agree that the results we have seen so far are *some* evidence in favour of a culturally mediated persistent effect of slavery on trust. The high degree of spatial autocorrelation makes me doubt the robustness of the results. But not enough to disbelieve them.

---

<sup>65</sup> See [table summary](#)

<sup>66</sup> I obtain "beta = -0.03 (0.04) [0.04], adjusted p = 4.33e-01, n=12024" after removing respondents from Benin, Ghana and Nigeria.

[https://github.com/Jsevillamol/persistence/blob/master/scripts/2c-nunn\\_wantchekon\\_2011-IV-ex-Benin.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/2c-nunn_wantchekon_2011-IV-ex-Benin.R)

# On the Origins of Gender Roles: Women and the Plough

by Alberto F. Alesina, Paola Giuliano and Nathan Nunn

## Table summary

<b>Result of replication</b>	Insignificant correlation between ancestral plough usage and female labor force participation amongst US american women children-of-immigrants	
<b>Main statistical method</b>	Children-of-immigrants OLS	
<b>Standardized <math>\beta \approx -0.01</math></b> (0.01) [0.03]	<b>Adjusted p-value <math>\approx 93\%</math></b>	<b>n = 79697</b> <b>mother birthplace countries = 126</b>
<b>Expected <math>\beta \approx 0.069^{67}</math></b>	<b># of hypothesis = 5<sup>68</sup></b>	<b>Critical hypothesis number = N/A</b>
<b>Power <math>\approx N/A</math></b>	<b>Type-S error rate <math>\approx N/A</math></b>	<b>Exaggeration ratio <math>\approx N/A</math></b>
<b>Moran's Z <math>\approx 0.26</math></b> <b>p-value <math>\approx 0.79</math></b>		<b>Persistence span <math>\approx 1800\sim 1950</math> to 2000<sup>69</sup></b>
<b>Short commentary</b>	Through the paper, a wider range of gender outcomes are studied and more dramatic results are reported. However I find them unsatisfying on various grounds. Most notably, I argue that income is a collider and controlling for it biases the results. Also the results fare poorly against MHT correction.	
<b>Reproduction details</b>	We reproduce the analysis of USA children-of-immigrants, regressing labour participation for females on mother's country of origin average ancestral usage of the plough (table IX column 2)	

<sup>67</sup>These results from the geometric mean of the standardised effect sizes found in table 3 of (Cipollone et al., 2014)

<sup>68</sup> As there are five outcomes being studied: female labour force participation rate, female firm ownership, women participation in politics, and two measures of gender attitudes

<sup>69</sup> The 1800 to 1950 period covers 95% of the observations (Kirby et al., 2018). "Table 2 reports country-level OLS estimates. In columns 1 and 2, the dependent variable is a country's female labour force participation rate (FLFP) in 2000" (Alesina et al., 2013)

[https://github.com/Jsevillamol/persistence/blob/master/scripts/3d-alesina\\_giuliano\\_nunn\\_2013-USA\\_children\\_of\\_immigrants.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/3d-alesina_giuliano_nunn_2013-USA_children_of_immigrants.R)

## Review conclusion

Through this paper, the authors attempt to establish a link between ancestral plough usage and five measures of gender norms: female labour force participation, female firm ownership, female participation in politics, and surveyed gender attitudes with respect to laboral roles and leadership.

Where data is available, these measures are studied at the national and subnational level throughout the world, and in children of US-American and European immigrants.

While many parts of this analysis are suggestive, some others do not hold up to scrutiny.

The authors find that the correlation with female labour force participation does not robustly hold between countries when controlling for country fixed effects (table V) nor within some of the individual countries surveyed by the authors (table VI). Additionally, after adjusting for MHT the result stops being significant<sup>70</sup>

When the authors perform the study of female labour force participation in immigrants' children in the US (table IX), they find that the correlation significantly decreases with respect to the country-level and ethnic-level analysis. However, this result is only statistically significant when controlling for national average income<sup>71</sup> - which might be a collider. The result also does not withstand MHT correction<sup>72</sup>.

Female participation in politics is only significantly correlated with plough usage when controlling for income (tables III and IV) - and given that income might be a collider it is unclear if the control is appropriate. The subsequent IV analysis does not withstand MHT correction<sup>73</sup>.

The World Value Survey data lacks ethnicity information. Thus when the authors examined gender attitudes analysis based on this data (table V) they needed to construct their own proxy for ethnicity. By the author's own admission, this makes the analysis less conclusive,

---

<sup>70</sup> After the MHT adjustment I find "beta = -0.78 (0.37)" "Adjusted p = 0.162"

[https://github.com/Jsevillamol/persistence/blob/master/scripts/3b-alesina\\_giuliano\\_nunn\\_2013-country\\_level\\_IV.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/3b-alesina_giuliano_nunn_2013-country_level_IV.R)

<sup>71</sup> When adding the controls for income and income squared to those for the individuals and the historical country, I find beta = -0.03 (0.02) [0.03], unadjusted p = 0.047. If both controls for income are omitted, ceteris paribus, I find beta = -0.01 (0.01) [0.03], unadjusted p = 4.06e-01.

[https://github.com/Jsevillamol/persistence/blob/master/scripts/3d-alesina\\_giuliano\\_nunn\\_2013-USA\\_children\\_of\\_immigrants.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/3d-alesina_giuliano_nunn_2013-USA_children_of_immigrants.R)

<sup>72</sup> Controlling for income, the raw p-value is 0.047. The adjusted p-value is 0.212.

[https://github.com/Jsevillamol/persistence/blob/master/scripts/3d-alesina\\_giuliano\\_nunn\\_2013-USA\\_children\\_of\\_immigrants.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/3d-alesina_giuliano_nunn_2013-USA_children_of_immigrants.R)

<sup>73</sup> When controlling for continent fixed effects, and the full suite of historical and contemporary controls, the raw p-value is 0.393, and the MHT adjusted p value is 0.918

[https://github.com/Jsevillamol/persistence/blob/master/scripts/3b-alesina\\_giuliano\\_nunn\\_2013-country\\_level\\_IV.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/3b-alesina_giuliano_nunn_2013-country_level_IV.R)

both due to the additional degrees of freedom and possible biases introduced by the proxy. Furthermore, for one of the outcomes measured - gender attitudes about leadership- I find that introducing additional fixed effects for region, country and continent eliminates the significance of the latter<sup>74</sup>. The outcome of gender attitudes when faced with job scarcity withstand better my subsequent analysis, but only at a middling statistical significance level after correcting for MHT<sup>75</sup>.

In the IV analysis the study of gender attitudes based on immigrants' children in Europe (table X) authors find a statistically and economically significant effect. But I cannot reproduce their results. Instead, I find a statistically and economically significant positive correlation when I don't control for income, and a statistically and economically significant negative correlation when I control for income<sup>76</sup>. This suggests that income is either a crucial control or a collider - and in the absence of a reason to go either way we cannot trust this result.

The OLS analysis and IV analysis (tables III and VIII) indicate a significant causal negative effect of plough usage on female firm ownership. However I cannot replicate their finding<sup>77</sup>.

In the end, the conclusion with respect to the effect of historical plough usage on female firm ownership is unclear, especially given the relatively smaller attention female firm ownership receives in the authors' analysis with respect to female labour force participation.

One cross-cutting comment regarding the instrumental variables analysis for all outcomes, there is also a reasonable concern that the instrument of land suitability used by the authors is not appropriate, since they use a modern measure of land suitability which is presumably affected by the ongoing use of land (Drewry, 2006). It is unclear to me how much we should weigh this concern when evaluating the IV analysis.

By this point it is clear that I am not persuaded by all the points made by this paper. What do I take away?

If we believe in the suitability of the instrument, the natural experiment on female labour force participation is not robustly significant at the level reported by the authors - but it is still borderline significant at  $p \approx 5\%$ . On the other hand, the effect is far smaller and less

---

<sup>74</sup> For the correlation between ancestral plow usage and gender leadership attitudes outcome I find "beta = 0.08 (0.07)" [1] "unadjusted p = 0.244"  
[https://github.com/Jsevillamol/persistence/blob/master/scripts/3c-alesina\\_giuliano\\_nunn\\_2013-world\\_value\\_survey.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/3c-alesina_giuliano_nunn_2013-world_value_survey.R)

<sup>75</sup> beta = -0.06 (0.03)" [1] "p = 0.021" [1] "Adjusted p = 0.100"  
[https://github.com/Jsevillamol/persistence/blob/master/scripts/3c-alesina\\_giuliano\\_nunn\\_2013-world\\_value\\_survey.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/3c-alesina_giuliano_nunn_2013-world_value_survey.R)

<sup>76</sup> When I do not control for income I find "beta = 0.17 (0.02) [0.03], adjusted p = 0.00e+00" [1] "unadjusted p = 5.01e-21". If I control for income I find ""beta = -0.16 (0.03) [0.05], adjusted p = 1.47e-08"

<sup>77</sup> In the IV analysis I obtain "beta = -0.34 (0.50)" [1] "p = 0.501" [1] "Adjusted p = 0.969"  
[https://github.com/Jsevillamol/persistence/blob/master/scripts/3b-alesina\\_giuliano\\_nunn\\_2013-country\\_level\\_IV.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/3b-alesina_giuliano_nunn_2013-country_level_IV.R)

robust on children-of-immigrants. This is weak evidence suggesting a persistent effect, mainly mediated through non-cultural factors, which is an interesting finding on its own.

# The Church, intensive kinship, and global psychological variation

by Jonathan Schulz, Duman Bahrami-Rad, Jonathan Beauchamp, and Joseph Henrich

## Table summary

<b>Main result</b>	Association between mother's country of origin historical length of exposure to the Western Church and western psychological measures (high individualism, low conformity, high trust) in European children of immigrants	
<b>Main statistical method(s)</b>	OLS analysis on children of immigrant based on the length of exposure to the western , clustered at the country level	
<b>Standardised <math>\beta</math></b> $\approx$ 0.12 (0.01) [0.03]	<b>adjusted p</b> $\approx$ 1.29e-14	<b>n</b> = 36 regions 45196 respondents
<b>Expected effect size</b> $\approx$ 0.10 <sup>78</sup>	<b># of hypothesis tested</b> = 4	<b>Critical # of hypotheses</b> = 1.59e+13
<b>Power</b> $\approx$ 98%	<b>Type-S error rate</b> $\approx$ 0%	<b>Exaggeration ratio</b> $\approx$ 1.01
<b>Moran's Z</b> $\approx$ -0.30 <b>p-value</b> $\approx$ 0.77	<b>Adjusted R<sup>2</sup></b> $\approx$ 0.06	<b>Persistence span</b> = 500~1500 to 2014 <sup>79</sup>
<b>Short commentary</b>	The paper is devoid of quantitative causal analysis, and instead relies heavily on anthropological, psychological and historical theories. While the children of immigrant and country level analysis hold, my reproduction of the European level analysis shows small effect sizes.	

<sup>78</sup> Based on the pooled effect sizes from the natural experiments of the papers I review, as explained in [section 2.2](#)

<sup>79</sup> "Exposure to the mediaeval Western [...] and Eastern [...] Churches at the country level. The inset shows the Western Church exposure for regions within Europe based on the diffusion of bishoprics between 550 and 1500 CE [...]."

Data sources for outcome include: "R. Inglehart et al., Eds. 2014. World Values Survey: All Rounds - Country-Pooled Datafile 1981-2014. [...] European Social Survey Rounds 1 to 7 Data, Data file edition 1.0. NSD - Norwegian Centre for Research Data, Norway–DataArchive and distributor of ESS data for ESS ERIC (2014)." (Schulz et al., 2019)



	The analyses are surprisingly robust to spatial autocorrelation, despite the localised nature of the regressors.
<b>Reproduction details</b>	We reproduce the children-of-immigrant regression of a measure of individualism on mother's country of origin historical exposure to the Western Church, table 5 of the paper. <a href="https://github.com/Jsevillamol/persistence/blob/master/scripts/4c-schulz_et_al_2019-children_of_immigrant_europe.R">https://github.com/Jsevillamol/persistence/blob/master/scripts/4c-schulz_et_al_2019-children_of_immigrant_europe.R</a>

## Review conclusion

In this paper the authors study the hypothesis that the Western church played a causal role in dismantling kin-based institutions (eg cousin marriage, polygamy, etc) in Western society, leading to persistent peculiarities in psychological measures in westerners as measured by various psychological outcomes.

This relationship is studied with a battery of analysis, comparing outcomes between countries, between European regions and between European adult children-of-immigrants.

Through the paper, there is an acknowledged gap - the authors do not make a rigorous attempt at establishing causality via an instrumental variable analysis nor other methods. Thus it is really hard to back the claims of this paper on causal grounds, beyond what is warranted by its anthropological, psychological and historical theoretical basis, which is beyond my expertise to assess.

Thus, I treat this paper as an attempt at falsification. Consequently we should be cautious from updating too much in the direction of a causal effect of the western Church on western psychology with the dismantling of kin intensive institutions as the main mediator, as posed by the authors.

With that in mind, does the paper establish the associations expected by the underlying theory? I'd argue it does.

The country level analysis shows the expected associations. We find some traces of spatial autocorrelation in the kin institution intensity outcome regression, but not in the western church exposure analysis<sup>80</sup>. Then again, the country level analysis is the one more likely to be exposed to missing confounders, and thus the least informative.

Indeed, (Kelly, 2020) takes a closer look at the country-level results, and finds that the results are not significant after adding fixed effects for the World Bank Regions. We note

---

<sup>80</sup> In the country level western church exposure I find a Moran  $Z=-0.19$ ,  $p=0.85$ , while in the KII regression I find  $Z = 2.29$ ,  $p=2\%$   
[https://github.com/Jsevillamol/persistence/blob/master/scripts/4a-schulz\\_et\\_al\\_2019-country-level.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/4a-schulz_et_al_2019-country-level.R)

however that these fixed effects should not affect the European regions and European children-of-immigrant analyses.

The European region analysis of church exposure points also shows a correlation with individualism, albeit it is less statistically significant and the effect size is quite smaller compared to the country-level analysis<sup>81</sup>.

The children-of-immigrant analysis is a better control for external factors. And it shows a result midway between the results found in the country level and the European region level, finding a statistically significant and large correlation between Church Exposure, kin institution intensity and the chosen psych outcomes, despite a small effective sample size (with  $n=36$ )<sup>82</sup>. I also find no residual spatial autocorrelation.

I am confused about why the results at the European level are quite different. That being said, the lack of spatial autocorrelation and implicit control for confounders through the analysis of children-of-immigrants is indicative of a non-spurious, culturally mediated correlation.

Regardless, I find this paper lacking in causal grounds. There is not much to be done - a more rigorous causal analysis would require the identification of a valid instrument or a mediator where we can be reasonably sure is not too affected by missing confounders.

---

<sup>81</sup> In the country level analysis I find "beta = 0.49 (0.07)", "p = 0.00" - while in the European region analysis I find "beta = 0.03 (0.01)" [1] "p = 0.02"  
[https://github.com/Jsevillamol/persistence/blob/master/scripts/4b-schulz\\_et\\_al\\_2019-european\\_regions.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/4b-schulz_et_al_2019-european_regions.R)

<sup>82</sup> For example, for the correlation in children-of-immigrants between individualism and length of exposure to the western church of the mother's birthplace I find "beta = 0.12 (0.01)", "p = 0.00". Despite the small sample size, the analysis does not seem underpowered.  
[https://github.com/Jsevillamol/persistence/blob/master/scripts/4c-schulz\\_et\\_al\\_2019-children\\_of\\_immigrant\\_europe.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/4c-schulz_et_al_2019-children_of_immigrant_europe.R)

# Persecution perpetuated: The medieval origins of anti-semitic violence in nazi Germany

by Nico Voigtländer and Hans-Joachim Voth

## Table summary

<b>Result</b>	Correlation between antisemitic violence after the Black Death and 20th century antisemitic violence in the interwar period	
<b>Statistical method</b>	OLS	
<b>Standardised <math>\beta \approx 0.11</math> (0.04)</b> [0.08]	<b>Adjusted p-value <math>\approx 4.5\%</math></b>	<b>n=320</b>
<b>Expected <math>\beta \approx 0.10</math><sup>83</sup></b>	<b># of hypothesis = 6</b>	<b>Critical H # = 7</b>
<b>Power <math>\approx 0.39</math></b>	<b>Type-S error rate <math>\approx 0\%</math></b>	<b>Exaggeration ratio <math>\approx 1.60</math></b>
<b>Moran's Z <math>\approx 0.53</math></b> <b>Moran's p <math>\approx 0.59</math></b>	<b>R<sup>2</sup> <math>\approx 0.05</math></b>	<b>Persistence span = 1348~1350 to 1920~1945<sup>84</sup></b>
<b>Short commentary</b>	The paper finds a surprising long term correlation between antisemitic violence after the Black Death and 20th century antisemitic violence in the interwar period, but (1) the rest of the results provided in the paper are either statistically weak or suspect, (2) there is no attempt either theoretical or statistical at establishing causality and (3) I find the interpretation of the results obtained unconvincing	
<b>Reproduction details</b>	Here I reproduce the correlation between towns that participated in the Black Death pogroms and towns with incidents of interwar violence against Jews (table VI, column I)	

<sup>83</sup> Based on the pooled effect sizes from the natural experiments of the papers I review, as explained in [section 2.2](#)

<sup>84</sup> "We use data on anti-Semitism at two points in time – the medieval period, and the years 1920-1945."

"As our indicator for violence against Jews in the Middle Ages, we code whether there was a pogrom in 1348-50" (Voigtländer & Voth, 2012)

## Review conclusion

This is a very unusual paper, in both its goals and its methods.

In terms of its goals, it is not intending to establish a direct causal effect between the main exposure (violent acts against Jews following the Black Death in 1349) and the long term outcomes (several measures of antisemitism in the 20th century interwar period like violence against Jews, nationalist vote and jew deportations) - but rather argues that these events are symptom of an underlying, persistent antisemitic culture that varies at the city level.

The paper is also packed with innovative ideas for addressing the subtle points that might bias their analysis:

- To address a skewed noise distribution in the analysis of some variables they perform a Poisson max likelihood regression.
- To address a possible non-linear response, it brings to the table propensity matching.
- To address spatial autocorrelation, it performs a geographical matching analysis<sup>85</sup>.
- To distinguish whether their measures are capturing underlying antisemitism in cities or just a bundle of variables that happen to be correlated, they construct a new variable as the principal component of the standardization of these variables.
- To study the conditions under which cultural persistence occurs they introduce some interaction terms related to trade and prosperity to their analysis.

This paper is at its best at finding a surprising, statistically significant, well-powered, and not spatially spurious correlations between medieval antisemitic violence and 20th century interwar antisemitic violence, and arguably for vote of the NSDAP (the Nazi party) in 1928 as well<sup>86</sup>.

But I find the rest of the paper quite inconclusive.

The rest of the interwar antisemitic outcomes studied needed some statistical manipulation to show a significant correlation (namely the Poisson ML estimation) or are not statistically significant.

Previous cultural indicators also raise doubts : Jewish settlement patterns in the year 1933, violence aimed at Jews at 1800s (the Hep Hep riots) and before the Black Death are not shown to be significantly correlated with the Black Death pogroms (tables III and IX). The authors' explanation is that the underlying antisemitic culture only surfaces when shocks happen like the Black Death or the rise of nazism, but I find this unconvincing.

---

<sup>85</sup> To their credit, we find no traces of spatial autocorrelation as measured by the Moran statistic, which lends credibility to this approach

<sup>86</sup> Though the case is less clear cut for the election results due to higher spatial autocorrelation of residuals, with "Moran's Z = 17.02" [1] "Moran's p = 0.00". However, the results withstand geographical matching (see table VII, column 4).

[https://github.com/Jsevillamol/persistence/blob/master/scripts/5-voigthlander\\_and\\_voth\\_2012.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/5-voigthlander_and_voth_2012.R)

Yet another eyebrow-raising result is that despite the correlation of Black Death pogroms with NSDAP vote in 1928 the correlation vanishes with respect to vote in 1933 (table XI). The authors' explanation is that the NSDAP rise in 1928-1933 was not motivated by historically rooted antisemitism, unlike their earlier growth - again, I am not persuaded.

Also, the statistical significance of the results when regressing on the principal component of the 20th century antisemitic measures mostly show effect within the 5% threshold of significance (table VII). In fact, when adding some controls to the small sample where accurate information about mediaeval pogroms is available the result stops being significant (table VII, column 2).

Furthermore, I haven't looked in depth, but the interaction analysis is likely to be underpowered<sup>87</sup>. Thus I take the conclusions about how cultural persistence might be affected by trade, immigration and prosperity with a grain of salt.

In summary, I think this paper has found a surprising long term correlation between the Black Death pogroms and the interwar pogroms, but it does not provide a compelling explanation of this phenomena.

---

<sup>87</sup> Since including interaction terms greatly reduces the power of a sample (Gelman, 2018)

# Trade, Institutions, and Ethnic Tolerance: Evidence from South Asia

by Saumitra Jha

## Table summary

<b>Result</b>	Towns that were mediaeval harbours are found to have lower interethnic violence in the 1850-1950 period	
<b>Statistical method</b>	2SLS Instrument: presence of nearby natural harbours	
<b>Standardised <math>\beta</math></b> $\approx$ -0.16 (0.34) [0.56]	<b>p-value</b> $\approx$ 3.21e-01	<b>n</b> =248
<b>Expected <math>\beta</math></b> $\approx$ 0.10 <sup>88</sup>	<b>Hypotheses tested</b> = 2 <sup>89</sup>	<b>Critical hypotheses number</b> = N/A
<b>Power</b> $\approx$ N/A	<b>Type-S error rate</b> $\approx$ N/A	<b>Exaggeration ratio</b> $\approx$ N/A
<b>Moran's Z</b> $\approx$ N/A <b>Moran's p</b> $\approx$ N/A	<b>R<sup>2</sup></b> $\approx$ 0.42	<b>Persistence span</b> = 700~1700 to 1850~1950 <sup>90</sup>
<b>Short commentary</b>	The author finds a statistically significant correlation between mediaeval ports and lower interethnic violence, but their subsequent IV analysis does not show a statistically significant causal relationship.	
<b>Reproduction details</b>	We show the results from table 6, a 2SLS analysis relating whether a town is a mediaeval port to whether there is a historical record of interethnic violence between 1850-1950, with the presence of a natural harbour as the instrument.	

<sup>88</sup> Based on the pooled effect sizes from the natural experiments of the papers I review, as explained in [section 2.2](#)

<sup>89</sup> The outcomes studied include the number of riots and whether there was a riot at all

<sup>90</sup> "I provide evidence that medieval overseas trading ports, which served as the geographical focuses of these exogenous Hindu-Muslim complementarities and Muslim entry into trade, were not only relatively peaceful locations for Hindu-Muslim interaction in the medieval period, they were also five times less prone to Hindu-Muslim riots and around 25 percentage points less likely than otherwise similar towns to experience any religious riot between 1850 and 1950"

"By the beginning of the eighteenth century, the era of Muslim trade dominance in the Indian Ocean was over, and many medieval trading ports ceased to be commercially important. Thus, for close to a thousand years, intergroup complementarities existed between Hindus and Muslims in medieval trading" (Jha, 2013)

The results are not statistically significant, so the power / type-S / exaggeration ratio does not apply.  
The dataset for these results is not publicly available, so we could not reproduce them nor measure the spatial autocorrelation of residuals.

## Review conclusion

In this paper, the author studies the relation between mediaeval ports.

The main result of the paper, showcased in table 4, is a statistically significant correlation at  $p < 1\%$  between medieval ports and historical records of interethnic (Hindu-Muslim) violence in South Asia, with  $\beta = -0.17(0.05)^{91}$ . This is a large effect size compared to the other papers we have been studying, but the study does not appear to be too underpowered.

However, the author identifies an instrumental variable for determining whether a town will become a mediaeval harbour - the presence of a natural harbour. After running a 2SLS analysis, the statistical significance disappears, indicating that the identified correlation might not be causal.

Sadly, the dataset for this paper is not publicly available so we cannot rule out that the correlation reflects spatial autocorrelation bias.

The analysis of causation through the paper is supported by a theoretical framework of interethnic complementarities developed in a previous paper by the same author. This hypothesis postulates that the persistent effect was mediated by the creation of persistent institutions that promote cooperation between Muslims and Hindus in those places where the two ethnic groups provide complementary services to each other that they could not appropriate (eg, in mediaeval ports the muslims provided hindus with access to their trade network). The framework is illustrated in two historical cases - Surat and Ahmedabad.

I am cautious about my interpretation of the latter part of the paper where mediation is analysed and evidence is supposedly presented in favour of this theoretical framework, since I found it hard to understand.

However I am cautiously sceptical, and find it unsupported. The results rely on interaction analysis that is likely to be underpowered, the interpretation of the results is not immediately obvious and the results are not convincingly statistically robust - with p-values dancing between 1% and 10%.

The author also shows significant correlations between mediaeval ports and current attitudes towards vaccination, participation in trade unions and participation in financial entities. But

---

<sup>91</sup> To estimate the standard deviation on the “any riots” variable we combine the variances given in table 2 [REF]  $\sigma_y = 0.49$ . The standard deviation of the “mediaeval port” variable can be estimated as usual given the number of towns that are and are not mediaeval ports,  $\sigma_x = 0.31$ . This is not a perfect estimation since this takes into account observations omitted in the actual regression.

this analysis makes no effort to account for covariates, and may just reflect differences between coastal and landlocked towns.

In summary, the author finds a correlation between mediaeval ports and lower interethnic violence, but fails to show a causal relationship.



# Long-term persistence

by Luigi Guiso, Paola Sapienza and Luigi Zingales

## Table summary

<b>Main result</b>	Correlation between Italian cities that received a gold medal for heroic resistance against the fascist regime in 1943-1945 and their mediaeval independence status	
<b>Main statistical method(s)</b>	IV analysis Instrument: Presence of a bishop in the city in mediaeval times	
<b>Standardised <math>\beta</math></b> $\approx 0.11$ (0.02) [0.04]	<b>p-value</b> < 1%	<b>n</b> =5180
<b>Expected <math>\beta</math></b> $\approx 0.10$ <sup>92</sup>	<b>Hypothesis tested</b> = 5 <sup>93</sup>	<b>Critical number of hypothesis</b> = 5040
<b>Power</b> $\approx 62\%$	<b>Type-S error rate</b> $\approx 0\%$	<b>Exaggeration ratio</b> $\approx 1.01$
<b>Moran's Z</b> $\approx 1.75$ <b>Moran's p</b> $\approx 8\%$	<b>R<sup>2</sup></b> $\approx 0.32$	<b>Persistence span</b> = 1176 to 1943~1945 <sup>94</sup>
<b>Short commentary</b>	There is a likely causal correlation between mediaeval history of the cities and the likelihood to receive a gold medal for fascist resistance. I however find the analysis of modern measures of civic capital unsatisfactory, and I directly falsify the proposed mechanism for persistence via the establishment of a cultural trait of self efficacy.	
<b>Reproduction details</b>	Here I extend the results of the paper with an IV analysis on the gold medals regressed on the independence status of Italian cities in medieval times. The instrument used is the presence of a bishop in the city, as in the other IV analyses. <a href="https://github.com/Jsevillamol/persistence/blob/master/scripts/7b-guiso_et_al_2016-IV.R">https://github.com/Jsevillamol/persistence/blob/master/scripts/7b-guiso_et_al_2016-IV.R</a>	

<sup>92</sup> Based on the pooled effect sizes from the natural experiments of the papers I review, as explained in [section 2.2](#)

<sup>93</sup> Outcomes tested include number of local nonprofits, presence of a local organ donation organisation, propensity of local high schoolers to cheat in maths exams, local average psychological self-efficacy and whether the town received a gold medal

<sup>94</sup> "Our first definition of free cities includes all the cities that were independent in 1176." "Indicator = 1 if the city has been awarded a gold medal for the heroism of its population in fighting against the Nazi-Fascists in the 1943-1945 period." (Guiso et al., 2016)

## Review conclusion

In this paper, the authors study how Italian cities that became free city states in mediaeval times differ from their neighbours in terms of civic capital.

Civic capital is observed through four outcomes: 1) the number of nonprofits per capita in a city, 2) whether there is a local organ donation organisation, 3) prevalence of exam cheating among high schoolers and 4) whether the city received a gold medal for their heroic resistance of the fascist regime in 1943-1945.

The correlation with these outcomes is studied using OLS under some sets of controls. To identify whether the correlation is causal, the authors employ the presence of a city bishop at the before the period of formation of the free cities<sup>95</sup>.

The results of their analysis are, in my opinion, mixed.

I could not replicate the correlation between the number of nonprofits per capita and mediaeval free cities. My analysis also indicates a significant chance of spatial autocorrelation<sup>96</sup>, and that the effect size is quite small compared to other papers<sup>97</sup>,

For the outcome of high schoolers' propensity to cheat in a maths exam I don't find a significant correlation when reproducing the analysis<sup>98</sup> when reproducing the result. In the additional analysis conducted by the authors there are also several instances where this outcome is found to not be significantly correlated with mediaeval independence<sup>99</sup>, which indicates that this outcome is at best noisy and at worst not correlated with the exposure.

---

<sup>95</sup> The logic being that the bishop provided a mechanism for enforcing coordination and thus facilitating independence through the threat of excommunication. The instrument seems to be both strong and satisfy the exclusion restriction.

<sup>96</sup> For the correlation with the number of nonprofits the Moran statistic I get in the main analysis is  $Z \approx 25.67$ ,  $p < 1\%$ , and in the IV analysis is  $Z \approx 25.67$ ,  $p < 1\%$ .

[https://github.com/Jsevillamol/persistence/blob/master/scripts/7a-guiso\\_et\\_al\\_2016.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/7a-guiso_et_al_2016.R)  
[https://github.com/Jsevillamol/persistence/blob/master/scripts/7b-guiso\\_et\\_al\\_2016-IV.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/7b-guiso_et_al_2016-IV.R)

<sup>97</sup> For number of nonprofits in the main analysis I get a standardised coefficient of  $\beta \approx 0.02$  (0.02) [0.03], adjusted  $p \approx 0.68$ , and in the IV analysis is  $\beta \approx 0.04$  (0.03) [0.05], adjusted  $p \approx 0.69$

[https://github.com/Jsevillamol/persistence/blob/master/scripts/7a-guiso\\_et\\_al\\_2016.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/7a-guiso_et_al_2016.R)  
[https://github.com/Jsevillamol/persistence/blob/master/scripts/7b-guiso\\_et\\_al\\_2016-IV.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/7b-guiso_et_al_2016-IV.R)

<sup>98</sup> For cheating in the main analysis I get a standardised coefficient of  $\beta \approx 0.00$  (0.03),  $p \approx 1$  and in the IV analysis  $\beta \approx -0.08$  (0.05),  $p \approx 0.10$

[https://github.com/Jsevillamol/persistence/blob/master/scripts/7a-guiso\\_et\\_al\\_2016.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/7a-guiso_et_al_2016.R)  
[https://github.com/Jsevillamol/persistence/blob/master/scripts/7b-guiso\\_et\\_al\\_2016-IV.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/7b-guiso_et_al_2016-IV.R)

<sup>99</sup> For example, in in table 5 panel C column III we find that affiliation with Lombard League (that the authors argue is a proxy for a stronger degree of independence) is more strongly correlated with the cheating metric than being neutral in the League vs emperor Fredrick I conflict

The outcome of organ donation is more robustly correlated<sup>100</sup>, but I couldn't rule out spatial autocorrelation bias<sup>101</sup>.

Perhaps most surprisingly, while the authors focus the least on the gold medal outcome it is the most robust of all - its correlation with the mediaeval independence status of the cities is statistically significant, has a large effect size, robust, robust on the IV analysis, not spatially spurious and correctly powered.

Since this earlier measure is more robustly correlated than the other measures of civic capital this might indicate an elasticity in the cultural shift (shifts become less significant through time), that something happened between WWII and the 2000s that affected the variation or that the model of causation proposed in the paper is mistaken.

The authors also attempt in this paper to disentangle the causal mechanism by proposing that the relation is mediated by the establishment of a cultural trait of self-efficacy. I argue that this hypothesis is flawed - through a IV analysis using the bishop instrument preferred by the authors we can show that it is not causally related to the independence of the free cities<sup>102</sup>.

In summary, there is a likely causal correlation between mediaeval history of the cities and the likelihood to receive a gold medal for fascist resistance. This indicates a long term persistence mechanism.

However I find the analysis of modern measures of civic capital unsatisfactory, and I directly falsify the persistence mechanism of self-efficacy culture put forward by the authors.

## Acknowledgements

This paper has been authored by Jaime Sevilla of Aberdeen University.

---

<sup>100</sup> For the organ donation outcome I find  $\beta \approx 0.19(0.01)$ ,  $p < 1\%$  in the main analysis and  $\beta \approx 0.42(0.02)$ ,  $p < 1\%$  in the IV analysis

[https://github.com/Jsevillamol/persistence/blob/master/scripts/7a-guiso\\_et\\_al\\_2016.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/7a-guiso_et_al_2016.R)  
[https://github.com/Jsevillamol/persistence/blob/master/scripts/7b-guiso\\_et\\_al\\_2016-IV.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/7b-guiso_et_al_2016-IV.R)

<sup>101</sup> For the organ donation outcome I find a Moran  $Z \approx 7.81$ ,  $p < 1\%$  and for the IV analysis  $Z \approx 7.23$ ,  $p < 1\%$

[https://github.com/Jsevillamol/persistence/blob/master/scripts/7a-guiso\\_et\\_al\\_2016.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/7a-guiso_et_al_2016.R)  
[https://github.com/Jsevillamol/persistence/blob/master/scripts/7b-guiso\\_et\\_al\\_2016-IV.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/7b-guiso_et_al_2016-IV.R)

<sup>102</sup> When doing a 2SLS on self efficacy I find  $\beta = -0.01(0.05)$  [0.09], adjusted  $p = 1.00e+00$   
[https://github.com/Jsevillamol/persistence/blob/master/scripts/7c-guiso\\_et\\_al\\_2016-mediation.R](https://github.com/Jsevillamol/persistence/blob/master/scripts/7c-guiso_et_al_2016-mediation.R)

Thanks to David Rhys Bernard, Nathan Nunn, Aron Valinder, Luisa Rodriguez, William MacAskill, Florian Penner, Ehud Reiter, Max Marian Daniel, Faatima Osman, Pablo Villalobos, Ronja Lutz and Neel Nanda for feedback and constructive criticism.

Leticia García provided invaluable research assistance, including carefully reviewing all the analyses conducted in the paper. All remaining mistakes are solely my fault.

Pablo Villalobos, Ronja Lutz and Neel Nanda assisted me in acquiring and cleaning the data I needed for the analyses, for which I am incredibly grateful.

David Roodman and Julian Jamison provided a blind peer review of this paper.

This work has been supported by the Global Priorities Institute of Oxford University.

## Bibliography

Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). *Reversal of Fortune: Geography and Institutions in the Making of the Modern World Income Distribution* (Working Paper No. 8460; Working Paper Series). National Bureau of Economic Research.

<https://doi.org/10.3386/w8460>

Alesina, A., Giuliano, P., & Nunn, N. (2013). On the Origins of Gender Roles: Women and the Plough. *Quarterly Journal of Economics*, *128*(2), 469–530.

Altonji, J. G., Elder, T. E., & Taber, C. R. (2005). An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schooling. *The Journal of Human Resources*, *40*(4), 791–821. JSTOR.

Barret, M. (2020, February 12). *Common Structures of Bias*.

<https://cran.r-project.org/web/packages/ggdag/vignettes/bias-structures.html>

Borenstein, M. (Ed.). (2009). *Introduction to meta-analysis*. John Wiley & Sons.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.

<https://doi.org/10.1038/nrn3475>

Cipollone, A., Patacchini, E., & Vallanti, G. (2014). Female labour market participation in

- Europe: Novel evidence on trends and shaping factors. *IZA Journal of European Labor Studies*, 3(1), 18. <https://doi.org/10.1186/2193-9012-3-18>
- Dormann, C. F., McPherson, J. M., Araújo, M. B., Bivand, R., Bolliger, J., Carl, G., Davies, R. G., Hirzel, A., Jetz, W., Kissling, W. D., Kühn, I., Ohlemüller, R., Peres-Neto, P. R., Reineking, B., Schröder, B., Schurr, F. M., & Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography*, 30(5), 609–628. <https://doi.org/10.1111/j.2007.0906-7590.05171.x>
- Drewry, J. J. (2006). Natural recovery of soil physical properties from treading damage of pastoral soils in New Zealand and Australia: A review. *Agriculture, Ecosystems & Environment*, 114(2), 159–169. <https://doi.org/10.1016/j.agee.2005.11.028>
- Gelman, A. (2018, March 15). *You need 16 times the sample size to estimate an interaction than to estimate a main effect.*  
<https://statmodeling.stat.columbia.edu/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/>
- Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, 9(6), 641–651. <https://doi.org/10.1177/1745691614551642>
- Guiso, L., Sapienza, P., & Zingales, L. (2003). People's opium? Religion and economic attitudes. *Journal of Monetary Economics*, 50(1), 225–282. [https://doi.org/10.1016/S0304-3932\(02\)00202-7](https://doi.org/10.1016/S0304-3932(02)00202-7)
- Guiso, L., Sapienza, P., & Zingales, L. (2016). Long-Term Persistence. *Journal of the European Economic Association*, 14(6), 1401–1436. <https://doi.org/10.1111/jeea.12177>
- Ilter, C. (2017). WHAT ECONOMIC AND SOCIAL FACTORS AFFECT GDP PER CAPITA A STUDY ON 40 COUNTRIES. *Journal of Global Strategic Management*, 11(2), 51–62. <https://doi.org/10.20460/JGSM.2018.252>

- Jha, S. (2013). Trade, Institutions, and Ethnic Tolerance: Evidence from South Asia. *The American Political Science Review*, 107(4), 806–832. JSTOR.
- Kelly, M. (2019). *The Standard Errors of Persistence* (SSRN Scholarly Paper ID 3398303). Social Science Research Network. <https://doi.org/10.2139/ssrn.3398303>
- Kelly, M. (2020). *Understanding Persistence* (CEPR Discussion Paper No. 15246). C.E.P.R. Discussion Papers. <https://econpapers.repec.org/paper/cprceprdp/15246.htm>
- Kirby, K., Greenhill, S., & Forkel, R. (Eds.). (2018). *Ethnographic Atlas*. Max Planck Institute for the Science of Human History. <https://d-place.org/contributions/EA>
- M.Sc.<sup>1</sup>, M. H., Cuijpers<sup>2</sup>, P. D. P., Furukawa<sup>3</sup>, P. D. T. A., & Ebert<sup>2</sup>, A. P. D. D. D. (n.d.). 4.2 *Random-Effects-Model | Doing Meta-Analysis in R*. Retrieved 6 August 2020, from [https://bookdown.org/MathiasHarrer/Doing\\_Meta\\_Analysis\\_in\\_R/random.html](https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/random.html)
- Muthukrishna, M., & Henrich, J. (2016). Innovation in the collective brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1690), 20150192. <https://doi.org/10.1098/rstb.2015.0192>
- Nunn, N. (2008). The Long Term Effects of Africa's Slave Trades. *Quarterly Journal of Economics*, 123(1), 139–176.
- Nunn, N., & Wantchekon, L. (2011). The Slave Trade and the Origins of Mistrust in Africa. *American Economic Review*, 101(7), 3221–3252. <https://doi.org/10.1257/aer.101.7.3221>
- Roser, M., Ritchie, H., & Ortiz-Ospina, E. (2013). World Population Growth. *Our World in Data*. <https://ourworldindata.org/world-population-growth>
- Schulz, J. F., Bahrami-Rad, D., Beauchamp, J. P., & Henrich, J. (2019). The Church, intensive kinship, and global psychological variation. *Science*, 366(6466), eaau5141. <https://doi.org/10.1126/science.aau5141>
- Šidák, Z. (1967). Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *Journal of the American Statistical Association*, 62(318), 626–633.

<https://doi.org/10.1080/01621459.1967.10482935>

Tobler, W. R. (2005). *Movie Simulating Urban Growth in the Detroit Region* Author ( s ) :

/paper/Movie-Simulating-Urban-Growth-in-the-Detroit-Region-Tobler/ea5eefedd4

fa34b7de7448coc8e0822e9fdf956

Voigtländer, N., & Voth, H.-J. (2012). Persecution Perpetuated: The Medieval Origins of

Anti-Semitic Violence in Nazi Germany\*. *The Quarterly Journal of Economics*,

127(3), 1339–1392. <https://doi.org/10.1093/qje/qjs019>

## Appendix A: Explanation of the table summary

For each paper I reviewed I included a table summary of a replication attempt I made of the result I considered more relevant in the paper. This summary includes:

<b>Result</b>	Explanation of the result replicated
<b>Statistical method</b>	Statistical method I used to replicate the result
<b>Effect size</b>	Measured a <b>standardised <math>\beta</math></b> , with the standard error before multiple hypothesis testing (MHT) correction in parenthesis and after MHT correction in brackets. The MHT correction I use is derived in <a href="#">appendix B</a>  <i>Example:</i> $\beta=-0.5$ (0.1) [0.2] means that an increase of one standard deviation in the exposure is associated with a decrease of between 0.3 and 0.7 standard deviations in the outcome with ~70% confidence after MHT correction
<b>p-value</b>	Two-tailed statistical significance of the result versus a null effect. In essence, the p-value is the complement of the width of the confidence band that would contain an effect size of 0. I report the p values adjusted for multiple hypothesis testing using the Šidák correction (Šidák, 1967).
<b>Persistence span</b>	The historic window being studied, in years.  <i>Example:</i> a persistence span of 1600~1700 to 2003 indicates that the historical exposure was measured in the period of 1600 to 1700, and the modern outcome was measured in 2003
<b>n</b>	Sample size. Note that in many of the studies the observations are clustered in various strata, reducing the effective sample size.

<b>R<sup>2</sup></b>	Percentage of the variation in the outcome explained by the regressors, adjusted for degrees of freedom.
<b>Moran's Z &amp; Moran's p</b>	<p>The most pervasive mistake we will be looking out for is spatial autocorrelation bias: when the exposure and outcome values cluster together geographically, they are subject to a spurious spatial correlation.</p> <p>In particular, we need to watch out for the residuals of our linear regressions being spatially correlated, as that violates the Gauss-Markov assumption of uncorrelation.</p> <p>Following (Kelly, 2019) we will use the Moran I statistic to measure spatial correlation of residuals. We report the Moran statistics normalised as a Z-score, and its associated p-value vs the null hypothesis of no correlation.</p> <p>A high Z score is no proof that spatial correlation bias is at play, but it is a sign for caution.</p> <p>Note that unlike Kelly, we use weight matrices that correspond to the inverse distance between the centroids of observations (be it countries, districts, etc), so our results differ slightly.</p> <p>When many observations are taken from the same location, we average the residuals before computing the Moran statistic.</p> <p>Note that spatial autocorrelation may happen even if the exposure (E) and outcome (O) are surveyed in physically distinct places, as long as they share a similar spatial pattern. When this happens, measuring the residual's Moran with respect to either spatial distribution should help us discern the autocorrelation, as the spurious correlation appears via a link of E &lt;- ESL &lt;- SAC -&gt; OSL -&gt; O (SL stands for Spatial Location).</p>
<b>Expected effect size</b>	Standardised effect size that we expect to get on average, based on priors. I either estimate this based on previous literature or using the pooled effect size of the papers. This is used as an input to compute the power, type-S error rate and exaggeration ratio, see below.
<b>Power, type-S error rate and exaggeration ratio</b>	<p>I follow (Gelman &amp; Carlin, 2014) to compute the probability that a hypothetical replication would find statistically significant results (power), the probability that the sign of the estimate is mistaken (type-S error rate) and the expected exaggeration factor due to p-value selection bias (exaggeration ratio).</p> <p>I compute these statistics over the standardised effect size we attempt to replicate.</p>



	Note that these calculations require us to make a judgement call about the a priori expected effect size. I will try to ground this in previous literature where possible.
<b>Critical hypotheses number</b>	<p>The critical number of hypotheses that would reduce the p value below 5%, after adjusting for the Šidák multiple hypothesis testing correction (Šidák, 1967).</p> <p>Intuitively, this corresponds to a lower bound on the amount of cherry picking the author would have had to engage in to obtain a significant result from an insignificant relation.</p>
<b>Short commentary</b>	A brief summary of what I find most compelling and most doubtful of the analysis
<b>Reproduction details</b>	A link to the script I used to reproduce the results, indication on what results am I exactly reproducing and differences in methodology wrt the author

As each paper presents multiple analyses, I make a subjective choice on which coefficients and p-values to report. I will specify my choice on each summary and provide the code to reproduce my results when data is available. To choose the criteria I try to choose to report results that: 1) I could reproduce myself, 2) are conservative, 3) feel most relevant.

## Appendix B: Standard error adjustment for multiple hypothesis testing

We test  $m$  different hypotheses which produce results of the form  $r_i = \beta_i(\sigma_i)$ , and selectively report the one that has a higher absolute  $Z$  value  $Z_i = \beta_i / \sigma_i$ .

In a single hypothesis setting, the standard error  $\sigma_i$  defines a confidence interval with confidence level  $\alpha = \text{erf}(1/\sqrt{2}) \approx 68\%$ .

The Šidák correction (Šidák, 1967) adjusts the individual hypothesis confidence level to maintain the same level of overall confidence  $\alpha$ , using the formula  $p = 1 - (1-p)^m$ , or in terms of confidence  $\alpha' = \alpha^{1/m}$ .

The usual single hypothesis  $\alpha$ -confidence interval is defined as  $\beta_i \pm \sigma_i \sqrt{2} \text{erf}^{-1}(\alpha)$ .

Thus if we wish the usual confidence interval to remain within the confidence level  $\alpha$  in the multiple hypothesis testing setting we need to apply the correction  $\sigma'_i = \sigma_i \sqrt{2} \text{erf}^{-1}(\alpha')$ .

The resulting confidence intervals, corrected for multiple hypothesis testing, are of the form  $\beta_i \pm \sigma'_i \sqrt{2} \text{erf}^{-1}(\alpha)$ .