

Uprooted: Migration, Coercion, and the Roots of Social Connectedness*

Johan Fourie[†]

Abstract

How people migrated – not just whether they did – shapes its social legacy. Combining the Putterman-Weil World Migration Matrix with Meta’s Social Connectedness Index for 14,028 country pairs, I show that shared ancestry predicts bilateral Facebook friendships, surviving controls for genetics, language, religion, folklore, trade and migrant stocks. Within-country tests in US counties, Australasia and South Africa sharpen this; the contrast is by regime: voluntary settlement leaves a strong homeland-oriented trace, indenture an intermediate one, and coerced migration little or none. The conditions of migration determine whether demographic history becomes a durable social bond.

Keywords: social connectedness; migration; ancestry; gravity model; deep roots; Facebook

JEL codes: F22; N30; O10; Z13

*I thank Gabriel Brown, Etienne le Rossignol, Michael Muthukrishna, Nathan Nunn and Melanie Xue for helpful comments. This paper was produced with substantial assistance from large language models, principally Anthropic’s Claude, with OpenAI’s Codex used for independent code and reference checking. These tools helped to write and debug the analysis code, to formalise the cultural-transmission model, to prepare an initial draft, and to edit the manuscript. I have reviewed and verified all code, results and claims, and I alone am responsible for the paper, its arguments and any remaining errors. I record this openly: such tools are becoming part of how economics is done, and transparency about their use should be the norm. Cite this paper as: Fourie, Johan. 2026. “Uprooted: Migration, Coercion, and the Roots of Social Connectedness.” Working Paper, Department of Economics, Stellenbosch University.

[†]Department of Economics, Stellenbosch University. Email: johanf@sun.ac.za.

1 Introduction

New Zealand and the United Kingdom are 18,000 kilometres apart. Yet their Facebook Social Connectedness Index – a measure of the relative probability that two users in different countries are friends on the platform – is among the highest bilateral scores in the world. New Zealand and Papua New Guinea, by contrast, are neighbours separated by less than 4,000 kilometres, yet their social connectedness is unremarkable. Generations of British migration to New Zealand created an ancestral bond that persists in the digital age. The pattern is not unique to New Zealand. Figure 1 compares six countries’ social connectedness footprints. The settler colonies of the United States, New Zealand and South Africa show long-distance ties to ancestral homelands in Europe; non-settler countries such as Papua New Guinea, Nigeria and Uzbekistan show gravity-dominated patterns concentrated among neighbours. These non-settler footprints do display some secondary long-distance ties – Nigeria to the United Kingdom and the United States, Uzbekistan to Russia – but these run to large, populous economies and so are consistent with a gravity baseline in which destination “mass” matters, rather than reflecting the dense, homeland-oriented diaspora networks visible for the settler colonies. The visual contrast invites a more systematic question.

Social connectedness footprints: settler colonies vs non-settler countries

Left: settler colonies whose populations were reshaped by post-1500 migration. Right: non-settler countries.



Figure 1: Social connectedness footprints of six countries. Left column: settler colonies (United States, New Zealand, South Africa) whose populations were reshaped by early-modern migration. Right column: non-settler countries (Papua New Guinea, Nigeria, Uzbekistan). Darker shading indicates higher Facebook social connectedness with the focal country (shown in dark grey).

This paper asks whether such ancestral bonds predict the structure of modern social networks systematically. I combine the Putterman and Weil (2010) World Migration Matrix, which records

the share of each country’s year-2000 population descended from every origin country since 1500, with Meta’s Social Connectedness Index (SCI; Bailey et al., 2018), which measures bilateral Facebook friendships across 168 countries. The Putterman-Weil matrix was designed to study the deep determinants of economic development; the SCI provides an unprecedented window into the global structure of social ties through the revealed friendships of 2.9 billion users. I ask: do countries that share more ancestral populations maintain stronger digital social ties today?

At the country level, the answer is suggestively yes: in the benchmark log specification a one-standard-deviation increase in shared ancestry is associated with roughly a 10 percent increase in bilateral connectedness, conditional on distance, common language, colonial history and country fixed effects. In a horse race against genetic distance (Spolaore and Wacziarg, 2009), linguistic proximity, religious similarity, folklore traditions (Michalopoulos and Xue, 2021) and two centuries of trade, ancestry retains 58 percent of its baseline magnitude and stays highly significant, and it remains significant after controlling for contemporary UN migrant stocks – so the relationship is not reducible to current diasporas maintaining homeland friendships. The effect is concentrated in settler societies, the pattern visible in Figure 1.

The country-level result, however, is descriptive: ancestry is not randomly assigned, no single horse race resolves the channels through which it operates, and the magnitude is sensitive to estimator choice. The paper’s central claim therefore comes from a different place. Three subnational tests – in the United States, Australasia and South Africa – exploit variation in the *type* of historical migration. Voluntary settlement, indentured labour and coerced migration coexist within these settings, allowing a contrast that the country-level horse race cannot deliver. The result is not symmetric across the three regimes: voluntary movement leaves a strong homeland-oriented social trace, indenture an intermediate one, and coerced migration leaves no comparable trace. Whether persistence operates depends on whether migrants could keep families together, build institutions and reproduce community life. I read this through the lens of what I call the *social infrastructure* of migration – the family ties, diaspora institutions and cultural practices that can connect origin and destination across generations – but the evidence speaks for itself: the conditions of migration determine whether demographic history becomes a durable social bridge.

The first test exploits within-country variation in US counties. Counties with more Irish ancestry are more connected to Ireland on Facebook; counties with more Italian ancestry are more connected to Italy; and the same positive within-state pattern appears for all nine origin groups tested. Ireland remains one of the strongest cases, but the county estimates also reveal substantial effects for the United Kingdom, Germany, Mexico, India and the Philippines. The county evidence therefore supports homeland-specific persistence, rather than lining up cleanly with a simple concentrated-versus-diffuse ranking (Sequeira, Nunn and Qian, 2020; Abramitzky and Boustan, 2022). This finding parallels Burchardi, Chaney and Hassan (2019), who show that US county ancestry composition predicts bilateral foreign direct investment with ancestral homelands; I extend their logic from investment flows to social networks.

The second test compares two settler colonies of the Pacific that occupy opposite ends of the

free-coerced spectrum. New Zealand was settled by free migration: organised settlement schemes such as the Canterbury Association and the Free Church of Scotland recruited families and built durable institutions. The 1881 New Zealand Census records borough-level birthplace, and those demographic patterns predict modern Facebook friendships in a district-level regression ($\beta = 0.317$ across 29 GADM2 districts). Australia, by contrast, was the destination for 162,000 transported convicts between 1788 and 1868 – people moved against their will, often without knowing where they were going, and frequently torn from their families on arrival. Penal colonies (New South Wales, Tasmania) and free colonies (South Australia, founded 1836) sit side by side in the same modern data, allowing a within-Australasia contrast in the conditions of migration. The pooled regression points in the predicted direction: the free regime (NZ regions and the free Australian colonies) gives $\beta = 0.250$ ($p = 0.009$), while the penal regime (NSW and Tasmania) gives $\beta = -0.865$ ($p = 0.12$, identified from two colonies). The Australian sample is too small to reject any null on either an asymptotic or a permutation basis, so the AU comparison is best read as directionally consistent with the South African slavery results rather than as a free-standing test of Hypothesis 3.

The third test exploits the same logic at finer resolution within South Africa. Voluntary European settlement (Dutch from 1652, British from 1820, French Huguenots from the 1680s) coexists in the same modern districts with indentured migration from India to Natal (1860–1911) and with slavery in the Cape Colony (from Mozambique, Madagascar, Indonesia and India). Voluntary settlement strongly predicts connections with European homelands ($\beta = 0.578$, $p < 0.001$), a modern district-level Indian share predicts stronger connections with India ($\beta = 0.078$, $p = 0.002$), and the direct slave-origin exercise yields no comparable positive coefficient ($\beta = 0.010$, $p = 0.22$). That slave-origin test is identified primarily from Mozambique-linked variation, so I read it as evidence against a robust positive slave-origin tie rather than as a precise average across all origin regions. A kitchen-sink specification leaves the slave-origin coefficient a small, fragile null once a mechanical pooling artifact is corrected (Section 8.3); it is in no case the positive tie that voluntary settlement produces. Trans-Atlantic and intra-American slave-trade corridors point in the same direction: forced-migration corridors do not predict modern homeland-oriented connectedness. The South African contrast, the Atlantic extension, and the Australian convict comparison all converge on a single claim: durable cross-border ties depend on migrants’ ability to maintain families, institutions and homeland links across generations.

The paper makes four contributions. First, the paper contributes to the economics of migration (Sequeira, Nunn and Qian, 2020; Burchardi, Chaney and Hassan, 2019; Tabellini, 2020; Blumentstock, Chi and Tan, 2025; Fourie and Santana-Gallego, 2013). Much of that literature studies how migration reshapes destination economies, politics and network value. Burchardi, Chaney and Hassan (2019), for example, show that US county ancestry predicts bilateral foreign direct investment with ancestral homelands, while Fourie and Santana-Gallego (2013) show that historical migration predicts tourism through ethnic reunion and cultural affinity. My contribution is to show that the same demographic history predicts a broader relational outcome: the geography of everyday social connectedness itself. That extends the migration literature from concrete flows such as investment,

tourism or contemporary migrant networks to the underlying cross-border social ties that can sustain those exchanges over the long run – and that the conditions under which migration occurred determine whether those ties form at all.

Second, it contributes to the economics of slavery, coercion and unfree labour (Acemoglu and Wolitzky, 2011; Nunn, 2008; Whatley, 2022; Wright, 2022). In Acemoglu and Wolitzky, coercion is about how labour can be extracted when freedom is denied; in Nunn and Whatley, the slave trades are shown to have devastating long-run effects on African development and institutions. This paper asks a different but complementary question: what kind of cross-border social legacy did coerced migration fail to create? The South African comparison shows that voluntary settlement, indenture and slavery generated sharply different patterns of modern homeland connectedness; the Atlantic and intra-American extension shows that the slavery null is not confined to one country; and the Australian convict comparison gives point estimates that are directionally consistent with this pattern, even though the five-colony sample is too small to reject the null at conventional levels. The contribution is to identify what coerced migration failed to create: durable positive origin-country ties of the sort visible after free movement.

Third, it contributes to the economics of cultural transmission and change (Bisin and Verdier, 2001, 2023; Giuliano and Nunn, 2021). In the classic Bisin-Verdier framework, persistence depends on what families and communities can transmit across generations. This paper brings that logic to an international-network outcome. The core claim is that migration can create a stock of *social infrastructure* – family ties, local institutions, community practices and homeland-oriented memory – that outlives the migration episode itself, and which persists only when migrants have the agency to build and reproduce it. The US, Australasia and South Africa results all point in the same direction: persistence is strongest where migrants could reproduce community life, and weakest where coercion disrupted exactly those channels.

Fourth, the paper contributes to the deep-roots literature and to the debate over persistence in economic history (Putterman and Weil, 2010; Spolaore and Wacziarg, 2009, 2018, 2013; Bazzi, Fiszbein and Gebresilas, 2020; Becker et al., 2016; Valencia Caicedo, 2019; Conley and Kelly, 2025; Arroyo Abad and Maurer, 2021). Putterman and Weil turned ancestry into a central object in development economics, and Spolaore and Wacziarg showed how deep historical distance can shape diffusion and development. I add a different outcome to that agenda – ancestry predicts present-day social connectedness directly, not only income, institutions or barriers to diffusion – and a different design. Rather than relying only on broad cross-country gradients of the sort questioned by Conley and Kelly (2025) and Arroyo Abad and Maurer (2021), the paper leans on historically specific within-country designs. The Australasian and South African exercises are especially useful here: within a single national setting, historical settlement patterns still predict modern ties to the homeland, and the type-of-migration contrast offers the kind of mechanism evidence that pure cross-country persistence regressions cannot.

Section 2 lays out a simple cultural-transmission model that grounds the voluntary-forced distinction. Sections 3–4 describe the data and the empirical strategy. Section 5 reports the country-

level results, sections 6–8 the three subnational tests (US counties, Australasia, South Africa), section 9 the robustness checks, and section 10 discusses the findings and concludes.

2 Theoretical Framework: The Social Infrastructure of Migration

Why should ancestral bonds – the demographic legacy of migrations that occurred centuries ago – predict the structure of 21st-century social networks? And why should this depend on whether the migration was voluntary or coerced?

I argue that migration creates what I call *social infrastructure*: the family connections, diaspora institutions, cultural practices and community organisations that link origin and destination across generations. The object of interest is therefore a stock of persistent social ties and informal institutions, not cultural values in the broad sense. Cultural transmission matters because families and communities help reproduce those ties, but the outcome measured here is social connectedness. This infrastructure is built by migrants and transmitted to their descendants through the mechanisms emphasised by the cultural transmission literature (Bisin and Verdier, 2023; Giuliano and Nunn, 2021). But – and this is the central theoretical claim – its construction requires *agency*: the freedom to settle among co-ethnics, to build homeland-style institutions, and to transmit homeland-oriented identity to children. Forced migration, by design, denies this agency.

I formalise this argument with a simple model of intergenerational transmission and then derive three testable hypotheses.

A model of homeland attachment

Consider a homeland-oriented attachment – the disposition and opportunity to maintain active social connections to the ancestral origin country. Let $q_t \in [0, 1]$ denote the share of the diaspora population carrying this attachment in generation t . Following Cavalli-Sforza and Feldman (1981), attachment is transmitted through two channels: *vertical* transmission from parent to child, and *oblique* transmission from the wider community.

Let $d \in [0, 1]$ denote the rate of vertical transmission – the probability that a parent who possesses homeland attachment successfully transmits it to a child through family socialisation (stories of the old country, maintained family ties, cultural practices in the household). With probability $(1 - d)$, vertical transmission from an attached parent fails, and the child’s trait is then determined by the community: with probability q_t the child encounters homeland-attached role models, institutions or peers, and adopts the trait through oblique socialisation. Children of non-attached parents do not receive the trait vertically, but may still acquire it through the same oblique channel. Aggregating across both types of parents gives the baseline transmission law

$$q_{t+1} = q_t [d + (1 - d)q_t] + (1 - q_t)q_t = q_t + dq_t(1 - q_t). \quad (1)$$

This is the Cavalli-Sforza and Feldman (1981) direct-plus-frequency-dependent model applied to

a binary trait at the population level. On its own, equation (1) is a persistence benchmark: for any $d > 0$ and any $q_0 \in (0, 1)$, the sequence converges monotonically to $q^* = 1$, while $q^* = 0$ remains an absorbing state if the trait is absent initially. That monotonicity is too strong for diaspora settings in which homeland attachment can erode through assimilation, intermarriage and generational decay. I therefore add a reduced-form attrition term $\alpha \in [0, 1]$ that captures the fraction of attached individuals who lose the trait between generations:

$$q_{t+1} = q_t + dq_t(1 - q_t) - \alpha q_t. \quad (2)$$

Under this extension, homeland attachment can either rise or fall depending on the balance between transmission and attrition. For any interior q_t , higher d raises finite-horizon persistence and higher α lowers it; the effect of a higher initial share is non-monotonic, because $\partial q_{t+1}/\partial q_t = 1 + d - \alpha - 2dq_t$ is positive only while $q_t < (1 + d - \alpha)/(2d)$ and turns negative once the community share is already large. When $\alpha \geq d$, the only stable steady state is disappearance; when $0 < \alpha < d$, the process converges to an interior steady state below one. The comparative statics are therefore no longer mechanical restatements of the initial condition: durable attachment requires transmission to outrun assimilation.

This stylised model remains deliberately reduced form, but with the attrition term it is at least capable of representing both persistence and erosion. The parameters should be read as containers for several concrete margins: family continuity, geographic clustering, institutional reproduction, social status, legal constraints, transport and communication costs, and active suppression of origin practices. The empirical designs do not separately identify each component. Their value is instead comparative: different migration regimes bundled these margins in sharply different ways. Three regimes of migration generate three predictions.

Voluntary migration. Voluntary migrants exercise agency in two dimensions. First, they maintain family structures, allowing $d > 0$: parents tell children about the homeland, maintain cross-border family ties, and celebrate cultural traditions. Second, they cluster geographically among co-ethnics (Bazzi, Fiszbein and Gebresilas, 2020; Abramitzky and Boustan, 2022), creating communities where q_0 is high. Ethnic associations, homeland-language churches, cultural organisations and community newspapers provide an institutional scaffold for oblique transmission. In the attrition-augmented model, voluntary migration is therefore characterised by relatively high d , high q_0 and comparatively low α : homeland attachment can persist because transmission and community reinforcement outrun assimilation. *Forced migration (slavery).* The slave trade systematically destroyed both transmission channels. Families were separated at the point of sale, severing the vertical channel: $d \rightarrow 0$. Slaveholders deliberately mixed ethnic groups to prevent collective resistance, driving the origin-specific local co-ethnic share – the community-level exposure q_0 that sustains oblique transmission – toward zero (Nunn, 2008). (Here q_0 is read as an origin-specific measure of local community density and exposure, consistent with its role in the transmission equation, rather than as a single initial prevalence of homeland attachment.) Cultural practices were suppressed and institutional formation prohibited by law, eliminating the community-level

infrastructure that supports oblique transmission and raising attrition pressure α . With low d , low q_0 and high α , homeland attachment decays rather than persists. This prediction applies most directly to *direct* measures of slave-origin ancestry. Broad modern racial categories, such as the Coloured population in the Cape, are much noisier proxies because they also pool Khoesan, African, European and mixed-origin histories.

Indentured labour. Indian labourers brought to South Africa’s KwaZulu-Natal province between 1860 and 1911 occupied an intermediate position. They had contracts rather than chains; they knew their specific origin; many maintained cultural practices; and many stayed voluntarily after their indenture ended. This implies $0 < d_{\text{indenture}} < d_{\text{voluntary}}$, $q_0 > 0$ but lower than for self-selected voluntary communities, and attrition pressure α below the slavery case but above the voluntary-settlement case. The model predicts an intermediate level of persistence: positive but weaker than for voluntary settlers.

Three hypotheses

The three regimes generate three testable predictions. *Hypothesis 1 (Ancestry and homeland connectedness):* contemporary ancestry composition predicts social connectedness with the corresponding homeland, and more concentrated settlement may strengthen that relationship by raising the density of co-ethnic institutions and oblique exposure. Under the attrition dynamics, $\partial q_1 / \partial q_0 = 1 + d - \alpha - 2d q_0$ is positive only while the community share is not already large, and under cultural substitution (Bisin and Verdier, 2001) parents invest more when the trait is locally rare; concentration is therefore best read as a directional summary of community formation and ethnic-institution density, not an exogenous treatment. *Hypothesis 2 (Institutional persistence):* the regional origins of *historical* settlement predict modern connectedness generations later, because organised 19th-century schemes – the Canterbury Association, the Free Church of Scotland in Otago – built churches, place names and cultural institutions that keep q_t high relative to attrition (Giuliano and Nunn, 2021; Bazzi, Fiszbein and Gebresilasse, 2020); identified within a single country, this is especially informative. *Hypothesis 3 (Freedom as a precondition):* voluntary settler ancestry predicts homeland connectedness while forced migration through slavery does not, with indenture intermediate – because slavery pushes d and q_0 down and α up, driving persistence toward zero rather than to a negative coefficient. The Cape Colony, where the same districts received voluntary Dutch, British and French settlers and enslaved people from Mozambique, Madagascar and Indonesia, is a natural laboratory for this prediction.

3 Data

3.1 Social Connectedness Index

The Social Connectedness Index, developed by Bailey et al. (2018), measures the relative probability that two Facebook users in different locations are friends on the platform. For locations i and j ,

the index is defined as

$$\text{SCI}_{ij} = \frac{\text{FB_Connections}_{ij}}{\text{FB_Users}_i \times \text{FB_Users}_j} \quad (3)$$

and is then scaled to have a maximum of 1,000,000. I use two versions. The country-level SCI covers 168 countries yielding 14,028 undirected pairs. The US county-to-country SCI covers 3,204 FIPS codes paired with 178 partner countries (570,312 observations). Both are from Meta’s 2021 data release, accessed via the Humanitarian Data Exchange.

Facebook had over 2.9 billion monthly active users in 2021. Bailey et al. (2018) validate the SCI against trade flows, patent citations, migration and disease transmission, and subsequent work uses it to study peer effects, trade and shock propagation (Bailey et al., 2022, 2021); Obradovich et al. (2022) show that Facebook interest data capture meaningful cultural variation across 225 countries.

The SCI has three relevant properties. It is strongly right-skewed – a few very high-SCI pairs (US-Canada, UK-New Zealand) and a long tail – which motivates the log dependent variable. It is available at fine geographic resolution, down to US counties and sub-national units globally, enabling the county analysis central to this paper. And it captures a broader notion of ties than migration or trade – family, study-abroad, professional and online connections – though the public release does not reveal whether a link reflects kinship, tourism, work or a one-time visit, nor does it contain interest or group data; I therefore interpret it as connection intensity, not tie content.

3.2 Putterman and Weil World Migration Matrix

The World Migration Matrix version 1.1 (Putterman and Weil, 2010) is a 172×172 matrix recording, for each country j , the share of its year-2000 population descended from each origin country i , tracing ancestry back to 1500. The matrix was constructed from historical demographic sources including census data, migration histories and population genetics, and represents the most comprehensive available accounting of the demographic legacy of five centuries of human migration.

I construct two ancestry variables. The first is $\text{anc_max}_{ij} = \max(\text{ancestry}_{ij}, \text{ancestry}_{ji})$, which takes the larger of the two directional ancestry shares. This bilateral summary is designed to capture the dominant directional demographic link: for almost all country pairs, non-zero ancestry runs mainly in one direction, so the maximum records that main connection without forcing a reciprocal structure that the data rarely contain. The second is a log transformation: $\text{anc_log}_{ij} = \log(1 + 1000 \times \text{anc_max}_{ij})$, which compresses the extreme sparsity of the matrix. Fully 91.5 percent of country pairs have zero shared ancestry, and the 99th percentile is only 0.041. The log transformation has a standard deviation of 0.693 compared with 0.028 for the raw variable, and yields the highest within- R^2 and significance level among all constructions tested (Appendix S). It is the preferred measure throughout, although I report alternative constructions and discuss the directional asymmetry explicitly in the robustness section.

3.3 Other sources

The country-level horse race draws on six families of bilateral measures: gravity controls from the CEPII Gravity Database (Head, Mayer and Ries, 2010); F_{ST} genetic distance (Spolaore and Wacziarg, 2009, 2018) and Out-of-Africa migratory distance (Ashraf and Galor, 2013); tree-weighted linguistic distance, religious distance, cultural distance from the World Values Survey, and folklore similarity (Spolaore and Wacziarg, 2016; Desmet and Wacziarg, 2021; Michalopoulos and Xue, 2021); bilateral trade flows in three periods from 1827 to 2014 (Fouquin and Hugot, 2016); contemporary bilateral migrant stocks (United Nations Department of Economic and Social Affairs, 2024); and trans-Atlantic and intra-American slave-trade corridors (SlaveVoyages Consortium, 2024). The US county-level analysis uses ACS 2018–2022 ancestry shares from Census Bureau tables B04006, B02015 and B03001 (US Census Bureau, 2022). The Australasian and South African subnational evidence draws on the 1881 New Zealand Census, the 1901 Australian Census via the Historical and Colonial Census Data Archive (Australian Bureau of Statistics et al., 2019), the 2022 South African Census (Statistics South Africa, 2024), the digitised Natal indentured records of Waetjen and Vahed (2014), the Cape Colony farm boundaries first published by Fourie (2022) and the slave emancipation dataset of Ekama et al. (2021).

Appendix M gives variable-by-variable definitions, coverage and construction details for each source, including the bilateral distance measures, the ACS extracts, the South African and Cape Colony datasets and the New Zealand and Australian census tables.

4 Empirical Strategy

4.1 Country-level specification

The baseline specification is a gravity model:

$$\log(\text{SCI}_{ij}) = \beta \cdot \text{anc_log}_{ij} + \mathbf{X}'_{ij}\boldsymbol{\gamma} + \alpha_i + \alpha_j + \varepsilon_{ij} \quad (4)$$

where anc_log_{ij} is the log ancestry measure defined above, $\mathbf{X}_{ij}$ includes log geodesic distance, contiguity, common official language and colonial ties, and α_i, α_j are country fixed effects. Because the data are stored as undirected pairs, standard errors are clustered on the two country identifiers that make up each dyad (the alphabetically ordered pair indices ‘i’ and ‘j’), following the multi-way gravity practice in Cameron and Miller (2015); Egger and Tarlea (2015). This accounts for correlation across pairs sharing a common country and approximately doubles standard errors relative to pair-level clustering. The coefficient β captures the partial association between shared ancestry and social connectedness, conditional on standard gravity determinants and all time-invariant country-level characteristics.

I build the specification sequentially, beginning with ancestry alone (column 1), adding distance (column 2), gravity controls (column 3), country fixed effects (column 4) and common religion (column 5). This sequence reveals how much of the raw ancestry-SCI correlation is absorbed by

geography, institutional history and unobserved country heterogeneity.

4.2 Horse race specification

The horse race augments equation (4) with competing bilateral channels:

$$\log(\text{SCI}_{ij}) = \beta \cdot \text{anc_log}_{ij} + \delta' \mathbf{Z}_{ij} + \mathbf{X}_{ij}' \boldsymbol{\gamma} + \alpha_i + \alpha_j + \varepsilon_{ij} \quad (5)$$

where $\mathbf{Z}_{ij}$ includes F_{ST} genetic distance, Ashraf-Galor migratory-distance proximity, linguistic distance, religious distance, folklore similarity, historical trade (three periods) and cultural distance. To keep the main table readable, Table 2 reports the richer and more standard bilateral channels, while Section 9.5 reports the Ashraf-Galor exercise separately because it is a coarser deep-history proxy that overlaps conceptually with F_{ST} and uses a slightly smaller sample. The key question is whether β remains positive and significant when competing channels are present. If it does, ancestry predicts bilateral connectedness beyond what is explained by those listed channels, though whether they are confounders or partial mediators of the ancestry effect is not something this regression design can resolve.

4.3 County-level specification

The county-level analysis exploits variation in ancestral composition across US counties within each origin-destination pair:

$$\log(\text{SCI}_{ck}) = \beta \cdot \log(\% \text{ancestry}_{ck}) + \delta \cdot \log(\text{pop}_c) + \theta_s + \varepsilon_c \quad (6)$$

where c indexes counties, k is the partner country, s is the state and θ_s are state fixed effects. Each regression is run separately for a single origin country k , and because the regressor is $\log(\% \text{ancestry})$, each is estimated on the counties that report a positive share of the relevant ancestry; the coefficient is therefore an intensive-margin elasticity, conditional on the ancestry being present, rather than an effect that also includes the zero-versus-positive margin. State fixed effects absorb all state-level confounders – port-of-entry effects, state immigration policies, regional economic conditions – so identification comes from within-state cross-county variation in ancestry composition. Standard errors are clustered at the state level to account for within-state spatial correlation in both ancestry patterns and social connectivity.

This is the paper’s tightest test. Within a country pair, all bilateral characteristics are constant. If counties with more Irish ancestry have higher SCI with Ireland – and counties with more Italian ancestry have higher SCI with Italy – this cannot be driven by any bilateral confounder. It can only be driven by county-level characteristics correlated with both ancestry composition and international social connections. State fixed effects absorb the most obvious candidates.

4.4 Identification discussion

The analysis is descriptive. Shared ancestry is not randomly assigned, and I make no causal claims. The identification toolkit is designed to assess the robustness of the descriptive association, not to establish causality.

The ancestry coefficient falls by 77 percent from the unconditional correlation to the full gravity-plus-fixed-effects specification but remains highly significant throughout, so the association is not entirely explained by geography or institutional similarity, even though the large reduction indicates substantial confounding; the residual coefficient’s interpretation depends on whether the controls are confounders or partial mediators. The remaining tools sharpen this: a permutation test places the coefficient 10.8–11.1 standard deviations above its null (Section 9); the horse race shows ancestry is not merely a proxy for another measured bilateral channel; and the county design holds all bilateral factors fixed, exploiting only within-state cross-county variation. Estimator choice matters too – log-OLS yields a positive, significant benchmark while PPML with country fixed effects is near zero (Santos Silva and Tenreyro, 2006) – so I treat the country-level coefficient as a descriptive benchmark and report full diagnostics in Section 9.

For completeness I report Oster (2019) bounds: the proportional selection coefficient is $\delta = 0.2$ ($R_{\max} = 0.906$), below the conventional threshold of 1, implying that unobservables only one-fifth as important as observables could eliminate the country-level coefficient. I therefore read it as a reason to weight the horse race, permutation, migrant-stock control and especially the subnational designs over the cross-country specification, not as supportive evidence.

4.5 South Africa district-level specification

The South Africa analysis tests whether ancestral composition predicts social connectedness within a single country. The baseline specification is:

$$\log(\text{SCI}_{dk}) = \beta \cdot \log(\% \text{White}_d) + \delta \cdot \log(\text{pop}_d) + \lambda \cdot \text{lang}_{dk} + \alpha_k + \varepsilon_{dk} \quad (7)$$

where d indexes the 52 GADM2 districts, k the partner, lang_{dk} a binary indicator for whether the district speaks a language historically linked to the partner, and α_k partner fixed effects. Distance is omitted from the European specifications because, with partner fixed effects, the within-South-Africa gradient in distance to European capitals is small relative to the common 9,000–10,500 km baseline; Appendix P shows the %White result is essentially unchanged when it is included. It is retained for non-European partners, where nearer, directionally varied partners generate informative within-partner cross-district variation (for India, distance varies along an east-west gradient correlated with the KwaZulu-Natal Indian concentration). Separate specifications pair $\log(\% \text{Indian}_d)$ with India and $\log(\% \text{Coloured}_d)$ with slave-origin countries.

For the Cape Colony subsample (14 districts), I add historical variables to the baseline:

$$\log(\text{SCI}_{dk}) = \beta_1 \cdot \log(\% \text{White}_d) + \beta_2 \cdot \log(\% \text{Surname}_d) + \lambda \cdot \text{lang}_{dk} + \delta \cdot \log(\text{pop}_d) + \alpha_k + \varepsilon_{dk} \quad (8)$$

where $\%Surname_d$ is a vector of district-level shares of slaveholders carrying historically Huguenot or 1820-settler British surnames, constructed from the slave emancipation records of Ekama et al. (2021). The forced-migration test uses the share of a district’s enslaved population originating from partner country k , matched via the same emancipation dataset. A farm-name robustness specification, classifying farms from the boundaries of Fourie (2022) by toponymic origin, is reported in Appendix N.

5 Country-Level Results

5.1 Baseline results

Table 1 presents the sequential OLS results. Without controls, the ancestry coefficient is 0.626 ($p < 0.001$). Adding log distance reduces it to 0.366, and adding the remaining gravity controls reduces it further to 0.232. Country fixed effects bring it to 0.144 ($p < 0.001$), a 77 percent reduction from the unconditional estimate. Adding common religion barely changes the coefficient (0.145), indicating that religion is a separate channel rather than a confounder. I treat column 4 as the benchmark log specification for the rest of the paper, while recognising below that the country-level magnitude is not invariant to estimator choice.

In the benchmark column 4, a one-standard-deviation increase in log ancestry (0.693) is associated with roughly a 10 percent increase in bilateral SCI ($0.693 \times 0.144 = 0.099$ in logs); these magnitudes are useful within the log-OLS framework but, as Section 9 shows, are not estimator-invariant.

Controlling for contemporary bilateral migrant stocks from the United Nations (United Nations Department of Economic and Social Affairs, 2024) halves the ancestry coefficient – from 0.145 to 0.073 – but it remains statistically significant ($p = 0.01$). The migrant stock variable itself is large and highly significant ($\beta = 0.154$, $p < 0.001$), confirming that contemporary diasporas contribute substantially to bilateral SCI. The survival of the ancestry coefficient suggests that part of the association reflects a stock of ties that persists beyond the current migrant generation – family links, institutional legacies and cultural connections that endure after the migrants who created them have passed – though I treat this as interpretation rather than a point-identified mechanism.

5.2 Horse race: ancestry versus competing bilateral channels

Table 2 presents the horse race. Column 1 repeats the baseline. Columns 2–7 add each competing channel individually. Columns 8–9 enter all channels simultaneously.

F_{ST} genetic distance is the strongest individual competitor, absorbing 35 percent of the ancestry effect (column 2: $\beta = 0.093$). This is expected: F_{ST} captures deep phylogenetic relatedness, and countries with shared ancestry tend to be genetically closer. But this reduction is not separately identifiable as “confounding” rather than mediation, because early-modern migration is itself one route through which populations became genetically more similar. The residual ancestry coefficient

Table 1: Baseline results: shared ancestry predicts bilateral social connectedness

	(1)	(2)	(3)	(4)	(5)
<i>Dep. var.: $\log(SCI_{ij})$</i>					
anc_log	0.626*** (0.058)	0.366*** (0.042)	0.232*** (0.038)	0.144*** (0.031)	0.145*** (0.032)
log distance		-1.453*** (0.059)	-1.394*** (0.054)	-1.468*** (0.052)	-1.420*** (0.056)
Contiguity			0.067 (0.150)	0.376*** (0.144)	0.302** (0.148)
Common language			1.561*** (0.100)	1.298*** (0.099)	1.259*** (0.096)
Colonial tie			0.916*** (0.194)	0.951*** (0.190)	0.930*** (0.177)
Common religion					0.510*** (0.111)
Country FE	No	No	No	Yes	Yes
R^2	0.053	0.414	0.505	0.697	0.704
Within R^2				0.560	0.568
N	14,028	14,028	14,028	14,026	13,039

Notes: Dependent variable: $\log(SCI_{ij})$. Standard errors clustered on the two country identifiers defining each undirected pair in parentheses. $\text{anc_log} = \log(1 + 1000 \times \max(\text{ancestry}_{ij}, \text{ancestry}_{ji}))$. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

should therefore be read as predictive power beyond F_{ST} , not as proof that the genetic channel is unimportant.

Folklore similarity is the second-strongest competitor (column 6: 28 percent reduction, $\beta = 0.104$). Shared mythological traditions proxy for deep cultural transmission through oral pathways that predate modern language families, religious boundaries and institutional structures. The folklore coefficient itself is large and highly significant (2.835^{***}), a novel finding that merits further investigation.

Linguistic distance absorbs 26 percent of the ancestry effect (column 3: $\beta = 0.107$). Language similarity is a natural mediator: countries that share ancestral populations often share languages, and shared language facilitates social connection. Again, the regression cannot distinguish whether the attenuation reflects omitted-variable control or the removal of a pathway through which ancestry operates. What remains is the component of the ancestry association not captured by this language measure.

Religious distance (column 4: 4 percent reduction) and historical trade (column 5: -4 percent, slight *increase*) have negligible confounding effects. The negative confounding from trade is notable: conditioning on historical trade slightly strengthens the ancestry coefficient, implying that trade and ancestry are partially offsetting channels rather than reinforcing ones. The negative pre-1870 trade coefficient (-0.035^{***}) may reflect the slave trade, where high historical trade is associated with low modern social connectedness.

Cultural distance (column 7) *increases* the ancestry coefficient to 0.160, but this reflects the restricted WVS sample (5,050 pairs) rather than genuine positive confounding. WVS countries are disproportionately settler colonies and wealthy democracies where the ancestry effect is strongest.

The kitchen-sink specification excluding cultural distance (column 8, $N = 11,323$) yields $\beta = 0.084$ ($p < 0.01$), a 41.5 percent reduction from baseline. The kitchen sink including cultural distance (column 9, $N = 4,369$) yields $\beta = 0.130$ ($p < 0.01$). In both cases the ancestry coefficient remains positive and precisely estimated in the preferred log-linear framework. Section 9 shows that the exact magnitude is sensitive to functional form, so the horse race should be read as evidence that ancestry carries incremental predictive power rather than as a definitive decomposition of mechanism.

Read in that way, the horse race speaks directly to the deep-roots literature: ancestry is not just a summary of genetic or cultural proximity, but a separate predictor of the social links through which diffusion and exchange may operate.

5.3 Where does ancestry matter? Settler colonies

Appendix V decomposes the global result by region. The ancestry coefficient is significant only for two subsamples: Americas $\times$ Rest ($\beta = 0.124$, $p < 0.01$; $N = 3,914$) and Europe ($\beta = 0.243$, $p < 0.01$; $N = 701$). It is insignificant for Africa ($\beta = 0.018$) and Asia ($\beta = 0.004$). The Americas-only subsample ($\beta = 0.224$) is insignificant, probably because the small sample ($N = 376$) lacks within-hemisphere variation.

Table 2: Horse race: ancestry versus competing bilateral channels

	(1) Baseline	(2) + F_{ST}	(3) +Ling.	(4) +Relig.	(5) +Trade	(6) +Folklore	(7) +Culture	(8) KS	(9) KS+C
<i>Dep. var.: log(SCI_{ij})</i>									
anc_log	.144*** (.031)	.093** (.029)	.107** (.032)	.137*** (.031)	.149*** (.031)	.104*** (.031)	.160*** (.036)	.084** (.030)	.130** (.040)
F_{ST}		-30.2*** (0.77)						-27.7*** (1.00)	-30.1*** (2.40)
Ling. dist.			-3.24*** (0.09)					-2.02*** (0.10)	-2.36*** (0.15)
Relig. dist.				-0.96*** (0.07)				-0.77*** (0.07)	-0.67*** (0.13)
Folklore sim.						2.84*** (0.10)		0.89*** (0.13)	-0.02 (0.25)
Cultural dist.							-12.5*** (0.87)		-5.85*** (0.94)
Trade controls	No	No	No	No	Yes	No	No	Yes	Yes
% reduction	—	35.0	25.7	4.3	-3.5	27.5	-11.6	41.5	9.4
Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Within R^2	.560	.612	.548	.567	.564	.587	.552	.606	.603
N	14,026	12,244	14,026	14,026	14,026	12,559	5,048	11,323	4,369

Notes: Dependent variable: $\log(SCI_{ij})$. All specifications include log distance, contiguity, colonial ties and country fixed effects. Columns without linguistic distance also include the binary common-language indicator; columns with linguistic distance use that continuous measure in its place. Standard errors clustered on the two country identifiers defining each undirected pair. Trade controls include log average bilateral trade in three periods: pre-1870, colonial era (1870–1945) and post-war (1946–2014). KS = kitchen sink excluding cultural distance; KS+C = kitchen sink including cultural distance. % reduction computed relative to baseline column (1). *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

The Americas $\times$ Rest subsample – the settler-metropole channel (US-UK, Brazil-Portugal, Argentina-Spain) – supplies most of the global identifying variation, though the estimated elasticity is actually larger within Europe (0.243 versus 0.124), reflecting large 19th- and 20th-century intra-European migration corridors. The null in Africa and Asia is consistent with the Putterman matrix capturing primarily early-modern European settler migration, which reshaped the Americas, Oceania and parts of Southern Africa far more than most African and Asian countries.

A complementary test interacts ancestry with a settler-pair dummy (equal to 1 if at least one country is an early-modern settler colony). The base ancestry effect for non-settler pairs is 0.123 ($p < 0.001$) and the interaction is 0.019 (insignificant). The total effect for settler pairs (0.142) is modestly larger, though the differential is not statistically significant. Settler pairs have a level effect of 0.272 ($p < 0.001$) on log SCI, indicating higher baseline connectedness.

The significant non-settler base effect (0.123) indicates that ancestry matters beyond the settler context: the settler-colony concentration reflects where most global variation lies, not a clean dichotomy in which ancestry matters only there.

5.4 Intensive margin

The extreme sparsity of the ancestry variable – 91.5 percent of pairs have zero shared ancestry – raises the question of whether the result is driven by the extensive margin (having any ancestry versus none) rather than the intensive margin (more versus less ancestry among positive pairs). Restricting the sample to the 1,199 pairs with non-zero shared ancestry, the coefficient increases to 0.244 ($p < 0.001$, $SE = 0.036$), larger than the full-sample estimate of 0.144. Among ancestrally linked pairs, the coefficient therefore remains positive and significant, which is consistent with a graded relationship. But the larger point estimate should not be over-interpreted: the positive-ancestry subsample is selected, so part of the increase may reflect its composition rather than a pure intensive-margin effect. The relationship is nonetheless genuinely graded – among ancestrally linked countries, more shared ancestry predicts more connectedness – and not merely the binary presence or absence of a tie.

6 Sub-National Evidence I: US Counties

6.1 Design

Hypothesis 1 predicts that contemporary ancestry composition predicts social connectedness, with the strongest effects for the most concentrated ancestral groups. The county-level analysis is designed to hold all bilateral characteristics constant. Within the US-Ireland pair, for example, language, colonial history, geographic distance and all other bilateral factors are fixed. The only variation is the ancestral composition of each county: some have more Irish ancestry and some have less, depending on historical patterns of Irish settlement.

I merge county-level SCI with Ireland to county-level Irish ancestry shares from the American Community Survey, and regress log SCI on log percent Irish ancestry, controlling for log population

and state fixed effects. State fixed effects absorb all state-level confounders, including port-of-entry effects (Irish immigrants concentrated in northeastern states), state-level economic conditions and immigration policies. Identification comes from within-state cross-county variation: among two counties in the same state, does the one with more Irish ancestry have higher SCI with Ireland?

These patterns reflect the legacy of the Age of Mass Migration documented by Bandiera, Rasul and Viarengo (2013) and Tabellini (2020), during which millions of Europeans settled across US counties in concentrations that still shape local demographics. I repeat this exercise for nine origin countries: the UK, Germany, Ireland, Italy, Mexico, Poland, France, India and the Philippines; Appendix B reports an expanded set that adds Nigerian, Ethiopian, Korean and Japanese origins.

6.2 Results

Table 3 reports the results. All nine origin countries show a positive and statistically significant association between county-level ancestry shares and social connectedness with the corresponding homeland. Ireland remains one of the strongest cases: a 10 percent increase in Irish ancestry in a county is associated with about a 3.5 percent increase in SCI with Ireland. Mexico shows the largest elasticity (0.577), followed by the UK (0.250), Germany (0.223), India (0.218), Italy (0.188), the Philippines (0.172), Poland (0.169) and France (0.115).

Table 3: County-level evidence: ancestry composition predicts bilateral social connectedness within US states

	Ireland	Italy	Poland	India	France	Philipp.	UK	Germany	Mexico
<i>Dep. var.: $\log(SCI_{c,k})$</i>									
$\log(\%ancestry)$	.371*** (.038)	.188*** (.020)	.169*** (.024)	.218*** (.019)	.115*** (.025)	.172*** (.019)	.250*** (.034)	.223*** (.029)	.577*** (.042)
$\log(pop)$	.176*** (.011)	.146*** (.008)	.219*** (.029)	.275*** (.021)	.159*** (.012)	.228*** (.022)	.229*** (.012)	.196*** (.008)	.101*** (.010)
State FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Within R^2	.288	.351	.237	.503	.291	.211	.381	.317	.390
N	3,140	3,121	3,007	1,893	3,127	2,367	3,145	3,150	3,133

Notes: Dependent variable: $\log(SCI_{c,k})$ where c is a US county and k is the partner country. Each column is a separate regression. County-origin shares use ACS 2018–2022 five-year extracts: B04006 for European ancestries, B02015 for Asian-origin groups, B03001 for Mexican origin, all divided by county total population (B01003). UK ancestry = English + Scottish. Standard errors clustered at the state level. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Mexico’s coefficient is especially large, consistent with dense contemporary cross-border social networks that remain strongly tied to local settlement patterns even after absorbing state fixed effects. India and the Philippines also show substantial elasticities, in line with active transnational diasporas. France is a different case: the French coefficient is more plausibly picking up older demographic legacies – especially French-Canadian, Cajun and earlier colonial settlement histories – than a recent skill-selected migration channel.

The near-universal pattern across very different diaspora groups – colonial-era European settlers, recent South Asian professionals and Southeast Asian immigrants – strongly suggests that ancestral composition has an independent association with modern digital social networks. The design is not

causal, and county-level confounders may remain, but the positive relationship appears across a wide range of ancestry-homeland pairs, which makes a single omitted explanation less likely. This is the paper’s cleanest test because the bilateral pair is fixed within each regression.

The county results are also robust to historical hard-infrastructure controls. In a separate county-level check, adding post-office density, Western Union presence, literacy, postal, telegraph and railway employment from Aneja and Xu (2024), and rail transport-access measures constructed from Donaldson and Hornbeck (2016) leaves all nine estimable ancestry coefficients positive and significant; the median coefficient retains 94 percent of its baseline magnitude.

The variation in elasticities across origin groups is informative. Irish and Italian coefficients are large, consistent with durable local ethnic institutions, but the sizeable British, German and Mexican estimates indicate that enduring homeland connectedness is not confined to the historically most concentrated European diasporas. Appendix C tests this concentration prediction of Hypothesis 1 directly with a Herfindahl index.

Placebo matrix. A natural concern is that the same county features driving Irish ancestry’s link to Ireland might also drive its link to any other partner. The full 9×9 placebo matrix (Appendix I) shows a clear average diagonal premium – mean diagonal coefficient 0.254 versus 0.079 off diagonal (paired $t = 3.22$, $p = 0.012$) – though 50 of 72 off-diagonal cells are also significant, because counties with many Irish-Americans also have many Italian-, German- and Polish-Americans, so any single ancestry share loads on the broader settlement pattern. To confirm the estimates are not merely loading on a gateway-county factor, Appendix O re-estimates each homeland regression jointly controlling for every other tracked ancestry and on a residualised “excess ancestry” measure; the focal coefficient stays positive and significant for all nine origins (for example, Ireland 0.267 and 0.189; Mexico 0.575 and 0.517).

A complementary piece of evidence comes from Americans who report “American” rather than a specific European ancestry on the Census – predominantly long-established British-descent populations in the South and Appalachia. Their county-level share is positively associated with SCI with the United Kingdom ($\beta = 0.061$, $p < 0.001$), and in a horse race with British ancestry, both variables are significant. The ancestry-SCI relationship therefore appears to operate through deep demographic roots even where subjective ethnic identification has dissolved (see Appendix D and Section 10 for further discussion).

These results also show that migration leaves durable origin-destination ties embedded in local population composition long after the original movers are gone. Burchardi, Chaney and Hassan (2019) document the analogous pattern for foreign direct investment; I confirm that the same demographic trace is visible in a different bilateral outcome – social networks – using a complementary identification strategy.

7 Sub-National Evidence II: Australasia – Free Settlement and Convict Transportation

This section uses two settler colonies of the South Pacific to test whether the conditions of migration – not just the fact of it – determine whether demographic history leaves a durable social trace. New Zealand was settled almost entirely through free migration. Australia, by contrast, received roughly 162,000 transported convicts between 1788 and 1868, alongside a far larger flow of free settlers. Penal colonies (New South Wales, Tasmania) and a deliberately free colony (South Australia, founded 1836) sit side by side in the same modern data, allowing a within-Australasia contrast that holds national institutions, language and the broader cultural environment broadly constant.

Hypothesis 2 predicts that the regional origins of historical settlement should predict modern social connectedness, even generations after the migration event, when migrants exercised the agency to build community life. Hypothesis 3 predicts that this persistence should weaken when migration was coerced – not only in the slavery case studied in Section 8, but also in the partly-coerced case of penal transportation. New Zealand provides the free-migration benchmark; Australia provides the partly-coerced contrast.

7.1 Organised settlement in New Zealand

New Zealand’s European settlement was concentrated in the mid-nineteenth century and organised along strikingly regional lines, each settlement relatively isolated and retaining a distinctive character (Olssen, 2025). Conceived along Edward Gibbon Wakefield’s lines of systematic colonisation, the New Zealand Company brought English settlers to Wellington, Wanganui, Nelson and New Plymouth in 1840–41 (Ballantyne, 2014). In 1848 the Otago Association, with the Free Church of Scotland, founded Otago to recruit Scots-Presbyterians; in 1850 the Canterbury Association established Christchurch as an explicitly Anglican English colony (Olssen, 2025). Auckland was more diverse, including a substantial Irish contingent and some British ex-convicts who crossed the Tasman – a small flow that slightly blurs the otherwise free character of New Zealand settlement – and 1860s West Coast gold drew Irish miners. Because locality of origin within Britain and Ireland is difficult to recover, I work with broad national birthplace categories (English, Scottish, Irish).

These organised settlement schemes created persistent regional demographic contrasts. The 1881 New Zealand Census – recording the birthplace of every non-Maori resident by borough – reveals the legacy: Otago’s borough populations were 20.4% Scottish-born (3.4 times the national average), Canterbury’s were 34.1% English-born (the highest of any region), and the West Coast’s were 13.6% Irish-born (also the highest). I aggregate 65 borough-level birthplace tabulations to 29 modern GADM2 districts for the preferred specification and to 13 GADM1 regions for descriptive figures.

7.2 Convict transportation to Australia

Australia is the natural counterpoint. Between 1788 and 1868 approximately 162,000 convicts were transported to the Australian penal colonies, the great majority from Britain and Ireland (Anderson and Maxwell-Stewart, 2014; Cowley et al., 2023; Nicholas, 1988). New South Wales received about 83,000 (1788–1850) and Van Diemen’s Land (modern Tasmania) some 73,500 (1803–1853); Western Australia received about 9,700 and roughly 3,000 “Pentonvillians” arrived at Port Phillip (modern Victoria) (Anderson and Maxwell-Stewart, 2014; Cowley et al., 2023). South Australia, by contrast, was founded in the 1830s explicitly as a free colony along Wakefield’s systematic-colonisation lines and received no convicts (Edmonds and Carey, 2017). Transportation was in any case dwarfed by free settlement: assisted schemes brought some 127,000 British and Irish settlers, and the non-Indigenous colonial population grew from around 70,000 in 1830 to more than a million by 1860 (Edmonds and Carey, 2017).

Convict transportation is not slavery – transported convicts kept their lives, served fixed sentences, knew where they came from and were eventually free in a colony where many chose to remain. But it shares the partly-coerced character that the paper’s central claim singles out. Convicts did not choose their destination, did not move with their families, and were assigned to masters and locations on arrival. The kinds of organised co-ethnic settlement that built the Canterbury Association’s English Anglican colony or Otago’s Free Church of Scotland community were absent. If the conditions of migration matter for the formation of durable cross-border social ties, then convict-heavy colonies should display a weaker linkage between UK-born share and modern UK-SCI than free-settled colonies.

I match colony-level convict-transportation totals from Anderson and Maxwell-Stewart (2014), Cowley et al. (2023) and Nicholas (1988) to the 1901 Australian Census birthplace tabulations from the Historical and Colonial Census Data Archive (Australian Bureau of Statistics et al., 2019), covering New South Wales, Victoria, Queensland, South Australia and Tasmania; Western Australia is excluded because comparable colony-level birthplace data are not available in the working sample. Queensland is treated as a free colony: its only penal episode was the secondary settlement at Moreton Bay (1824–1839), which received transportees reconvicted elsewhere in Australia rather than a primary transportation flow, so I do not count it as a convict-receiving colony. The convict-intensity measure is convict arrivals divided by the 1901 colony population, used both as a continuous covariate and to assign each colony to a regime category (*penal* for NSW and Tasmania; *free* for South Australia, Victoria and Queensland). Appendix M.5 documents the construction of the New Zealand and Australian variables.

7.3 Results

Because both samples are small, I report permutation p -values alongside asymptotic inference. The New Zealand evidence is reported in Panel A and B of Table 4; the pooled Australasia evidence appears in Appendix V.

New Zealand. I begin with the stricter specification. A stacked district $\times$ partner design combining seven tracked partner countries tests whether districts are especially connected to the *matched* historical homeland rather than simply being more internationally connected in general. Without fixed effects, the matched historical birthplace share is strongly positive ($\beta = 0.225$, $p < 0.001$; S1 in Table 4), but this largely reflects broad cross-partner differences in connectedness. Once partner fixed effects are introduced, the coefficient turns small and statistically insignificant ($\beta = -0.031$, $p = 0.184$; S2), and with both district and partner fixed effects it is essentially zero ($\beta = -0.001$, $p = 0.948$, permutation $p = 0.944$; S4). Adding GADM1 region fixed effects to the district-level UK-born specification likewise eliminates the coefficient ($\beta = -0.896$, $p = 0.394$; D4). New Zealand does not deliver a sharply local partner-specific result once one conditions on very demanding fixed effects. Appendix H pools these district-level regressions with Australia at the GADM2 level.

What *does* survive is the broader regional pattern. The share of the 1881 population born in the United Kingdom predicts 2021 Facebook SCI with the UK ($\beta = 0.317$, asymptotic $p < 0.001$, permutation $p = 0.143$, $N = 29$; specification D1 in Table 4). The Scottish-born and Irish-born shares are positive but imprecise ($\beta = 0.199$, $p = 0.106$; $\beta = 0.173$, $p = 0.290$). Read as descriptive evidence rather than as an identified treatment effect, these patterns fit the institutional persistence emphasised by Giuliano and Nunn (2021) and Bazzi, Fiszbein and Gebresilas (2020): the Canterbury Association, the Free Church of Scotland and the New Zealand Company created broad demographic and institutional contrasts whose echo remains visible in modern social connectedness.

Australia. The pooled NZ + AU regression delivers the directional pattern predicted by Hypothesis 3, but the Australian sample is too small to reject any null. Pooling 13 NZ regions and 5 AU colonies ($N = 18$) and regressing $\log(\text{SCI to GB})$ on UK-born share, log population and a country fixed effect, the semi-elasticity of UK-SCI to the UK-born share is $\beta = 0.217$ ($p = 0.065$; specification A1, Appendix V) – a one-percentage-point higher UK-born share is associated with roughly a 0.2 percent higher SCI with the United Kingdom, since the dependent variable is logged while the share enters in levels. The convict-share interaction in A2 is correctly signed ($\beta = -2.55$, $\text{SE} = 5.57$) but is not statistically distinguishable from zero on either inferential basis: the asymptotic p -value is 0.66 and the permutation p -value is 0.83. The free-versus-penal split (A3) reorganises the same five-colony variation into a categorical form: the free regime (NZ regions plus Queensland, Victoria and South Australia) yields $\beta = 0.250$ ($p = 0.009$), while the penal regime (NSW and Tasmania) yields $\beta = -0.865$ – but with $p = 0.117$, identified from only two colony observations. Tasmania’s convict share of 43% is roughly seven times NSW’s, so the interaction is mechanically dominated by Tasmania alone.

The AU point estimates are directionally consistent with Hypothesis 3, but the five-colony sample cannot test it. The Australasia exercise’s contribution is to show that adding the AU colonies to the New Zealand evidence does not overturn the free-settlement pattern. Appendix G reports the per-country region-level coefficients behind this pooled comparison.

Appendix U illustrates the New Zealand relationship – Otago and Southland, the Free Church

of Scotland settlements, stand out – and shows the full borough-level birthplace composition of each region in 1881.

8 Sub-National Evidence III: Voluntary and Forced Migration in South Africa

Hypothesis 3 predicts that voluntary settler ancestry should predict social connectedness with the homeland, while forced migration through slavery should not. South Africa provides a uniquely informative setting for this test. Five centuries of European settlement created a highly heterogeneous demographic landscape: Dutch settlement began in 1652, French Huguenots arrived in the 1680s, British settlers came in 1820, and indentured labourers from India were brought to KwaZulu-Natal from 1860. Crucially, the Cape Colony also imported tens of thousands of enslaved people from Mozambique, Madagascar, Indonesia, India and Malaysia. The coexistence of voluntary and forced migration in the same geographic setting allows a direct test of whether the conditions of migration determine the persistence of social connections.

8.1 Population composition and European social connectedness

Table 5 reports regressions for 52 GADM2 districts paired with 14 European partner countries (728 observations), using district-level population shares from the 2022 South African Census (Statistics South Africa, 2024). All specifications include partner fixed effects and $\log(\text{pop})$; distance is omitted from the European specifications because all South African districts are approximately equidistant from European capitals (9,000–10,500 km), leaving no meaningful within-partner variation. The share of the White population is a powerful predictor of social connectedness with Europe: $\beta = 0.578$ ($p < 0.001$, within $R^2 = 0.705$). This coefficient is stable across specifications. Adding a common language indicator – based on whether more than 5% of the district speaks Afrikaans (linked to the Netherlands) or English (linked to the UK) – reveals an independent language channel: $\beta_{\text{lang}} = 0.479$ ($p < 0.001$), raising within R^2 to 0.724.

The analysis provides three additional tests. First, the Indian/Asian population share predicts SCI with India ($\beta = 0.078$, $p = 0.002$), consistent with the 1860–1911 indentured migration; I retain distance to India as a control because the Indian population is concentrated on the east coast. Replacing the modern proxy with historical Natal indentured totals (Waetjen and Vahed, 2014) yields a smaller but still positive coefficient ($\beta = 0.015$, $p = 0.059$), though only 89,027 of 152,273 records are currently district-matched. Second, a placebo using six African neighbours shows that %White also predicts SCI with Africa ($\beta = 0.230$, $p < 0.001$) – whiter districts are more internationally connected in general – and adding district internet access does not eliminate it ($\beta = 0.243$), so the broader robustness exercises also condition on formal dwelling and urban shares (Appendix E). But a stacked specification pooling European and African partners reveals a highly significant %White $\times$ Europe interaction ($\beta = 0.348$, $p < 0.001$): the *differential* effect of %White on European versus African connectedness is large and robust, the most persuasive

evidence that the association is not merely generic cosmopolitanism. Third, a kitchen sink with all three groups shows White ($\beta = 0.404$), Coloured ($\beta = 0.125$) and Indian ($\beta = 0.040$) shares each positively associated with European connectedness (within $R^2 = 0.766$); I read the Indian term as a baseline connectedness gradient, with the placebo-stack comparisons isolating the White-European differential.

Figure 2 maps the contrast at the heart of Hypothesis 3: panel (a) shows strong spatial variation in connectedness with the UK – the primary homeland of voluntary settlers – while panel (b) shows no comparable variation in connectedness with Mozambique – the dominant origin of enslaved people. The per-partner scatter of $\log(\% \text{White})$ against $\log(\text{SCI})$ for the six largest European partners, and the White coefficient estimated separately for each European partner country – ranging from approximately 0.4 to 0.7 and all statistically significant – are shown in Appendix U.

To close off the concern that this merely captures generic cosmopolitan orientation, Appendix J extends the placebo logic to two regions with no historical South African tie – Asia (excluding India, Indonesia, Malaysia and the Philippines) and Latin America. The bare $\log(\% \text{White})$ coefficient is positive for both (Asia 0.265, LatAm 0.313), confirming a generic cosmopolitan axis, but the White $\times$ Europe interaction is large and consistent across all three comparisons (0.348 against Africa, 0.307 against Asia, 0.249 against Latin America): Europe sits roughly 0.25–0.35 above the shared cosmopolitanism baseline because it is the matched ancestral homeland. Within-district z-scoring of SCI reinforces the point (Appendix R): the European premium survives ($\% \text{White} \rightarrow \text{Europe} = 0.152$, $p < 0.001$) while the African placebo falls to marginal significance (0.028, $p = 0.09$) and the slave-origin coefficient is a clean null (-0.008 , $p = 0.33$).

8.2 Historical origins: the Cape Colony

The district-level analysis uses modern population groups as a crude proxy for ancestral composition. A sharper test exploits the slave emancipation records of Ekama et al. (2021), which list each enslaved individual’s owner surname. I link these surnames to a Huguenot surname dictionary and to a parsed 1820-settler surname list, building district-level shares of slaveholders carrying historically Huguenot or 1820-settler British surnames. Aggregating the 11 colonial Cape districts to modern GADM2 boundaries yields surname coverage for eight of the 14 Cape Colony districts in the analysis sample. The question is whether these historical settlement signatures add explanatory power beyond what modern demographics already capture.

Table 6 reports the result. The baseline White coefficient is strong in the Cape Colony subsample ($\beta = 0.785$, within $R^2 = 0.789$). Adding the 1820-settler owner-surname share reveals a significant independent channel ($\beta = 0.042$, $p < 0.001$), with the White coefficient falling to 0.698; the Huguenot owner-surname share enters similarly on its own ($\beta = 0.029$, $p < 0.001$). Entered jointly, the 1820-settler coefficient sharpens to 0.098 ($p < 0.001$) while the Huguenot coefficient flips to -0.045 ($p = 0.007$), so the two partly track the same coastal-versus-interior gradient and the British-settler signal is the more durable. The kitchen sink (C6), pooling all 21 EU and slave-origin partners, retains the pattern (1820-settler 0.088, Huguenot -0.046). The slave-origin coefficient

there is -0.030 ($p = 0.004$), but that is an artifact of a single $\log(\%White)$ slope across both partner groups; corrected, it collapses to a small, insignificant null (Section 8.3). Surname-based settler ancestry adds genuine explanatory power for European connectedness; forced-migration ancestry, properly identified, does not. Farm names (Fourie, 2022), classified by toponymic origin, give a directionally consistent but noisier signal (Appendix N).

8.3 Forced migration: slave origins and social connectedness

The Cape Colony presents a natural laboratory for testing whether forced migration creates the same durable social connections as voluntary migration. The same districts that received Dutch, British and French settlers also received enslaved people from Mozambique, Madagascar, Indonesia, India and Malaysia. If the social infrastructure created by migration depends on the conditions of that migration, then the demographic legacy of voluntary settlement should predict SCI with European homelands – but the demographic legacy of slavery should not predict SCI with slave-origin countries.

I test this two ways. First, at the full district level, the modern Coloured population share – a rough proxy for slave-descended populations – predicts SCI with slave-origin countries with a positive coefficient that does *not* disappear when internet access, formal dwelling and urban shares are added (β between 0.123 and 0.155; Appendix E). I therefore do not read it as support for Hypothesis 3: this broad category bundles slave-descended, Khoesan, African, European and mixed-origin histories and concentrates in the historically connected Western Cape (Appendix F explores a genetics-based decomposition). Second, and more sharply, I match the recorded origins of 914 enslaved individuals from the emancipation dataset of Ekama et al. (2021) to farms (Fourie, 2022), aggregating to districts. This matched variation is highly uneven – Mozambique appears in all 12 matched districts and accounts for about 80 percent of the sample – so the positive but insignificant baseline ($\beta = 0.010$, $p = 0.22$) should be read as a Mozambique-weighted null. In the pooled surname-led kitchen sink the slave-origin coefficient appears negative and significant ($\beta = -0.030$, $p = 0.004$), but this is a mechanical pooling artifact: imposing one $\log(\%White)$ slope across both partner groups forces large negative residuals on the slave-origin dyads. Interacting the European proxies with an EU-region dummy moves it to -0.003 , and re-estimating on slave-origin observations only ($N = 98$) gives -0.012 (SE 0.007, $p = 0.11$). Once the artifact is removed, the slave-origin tie is a small, fragile null, identified primarily from Mozambique.

Randomisation inference tells the same story. Permuting slave origins across districts 1,000 times, the baseline coefficient of 0.010 is indistinguishable from the null (two-sided permutation $p = 0.15$). The corrected slave-origin-only coefficient (-0.012) is only marginally distinguishable from its permutation null (two-sided permutation $p = 0.034$) and is far smaller in magnitude than the uncorrected pooled estimate, consistent with the artifact diagnosis above. This randomisation exercise does not create new cross-partner variation where the data are sparse, so it should be read as reinforcing the Mozambique-dominated null rather than as precise evidence for every minor slave-origin country. The null is nonetheless informative about magnitude: with a standard error

of 0.008, the design’s minimum detectable effect is roughly 0.023 at 80% power (Appendix Q), an order of magnitude below the voluntary-settler benchmark (%White $\rightarrow$ Europe = 0.578) and well below the indenture benchmark (%Indian $\rightarrow$ India = 0.078). The exercise can therefore rule out a homeland tie of the kind that free and semi-free migration produce.

Figure 3 summarises the contrast. Voluntary settler ancestry (White $\rightarrow$ Europe: $\beta = 0.578$) is associated with strong connections. Indentured migration (Indian $\rightarrow$ India: $\beta = 0.078$ using the modern population-share proxy; historical indentured totals: $\beta = 0.015$, $p = 0.059$) yields an intermediate association. Forced migration through slavery (slave origins $\rightarrow$ origin countries: $\beta = 0.010$, insignificant) creates no comparable positive relationship in the direct historical exercise, though that evidence is driven mainly by Mozambique because the matched slave sample is so concentrated there. This three-way comparison is consistent with the idea that durable cross-border ties depend on the conditions under which migration occurred. Enslaved people were violently separated from their origin communities, stripped of cultural practices and prevented from maintaining homeland ties. The new Indian indentured records sharpen the intermediate case, even if a fuller district-by-origin implementation remains the next step.

In that sense, the South African evidence speaks to the economics of slavery and coercion by identifying a missing legacy of unfree labour: coerced migration did not generate the durable homeland-oriented social ties observed after free and semi-free movement.

8.4 External validity: Atlantic and intra-American slave trades

The South African comparison is sharp but remains one country’s history. To test whether the logic travels, I return to the country level and use the Slave Voyages trans-Atlantic and intra-American databases to construct forced-migration corridors (SlaveVoyages Consortium, 2024). This is not a perfect replication – it uses corridor flows rather than descendant stocks, and the country pair is a noisier unit – so I treat it as a broader-scope test of Hypothesis 3. If coercion disrupted the reproduction of homeland ties, the major Atlantic slave-trade corridors should not display the durable positive SCI relationship that voluntary settlement does. This is distinct from the literature on slavery’s destructive effects on African development, institutions and trust (Nunn, 2008; Nunn and Wantchekon, 2011; Whatley, 2022); the question here is narrower – whether those corridors left durable *positive* origin-country social ties in today’s SCI.

All specifications follow the country-level framework of Section 4 (gravity controls, country fixed effects, two-way clustering); the appendix reports the full table. In an Africa-Americas sample of 265 mapped country pairs (263 enter the regression after gravity controls), a full-sample specification combining the presence and intensity of a slave corridor yields only a weak positive association ($\beta = 0.025$, $p = 0.10$), essentially unchanged when ancestry is added ($\beta = 0.026$) and roughly one-sixth the benchmark ancestry coefficient of 0.144 – consistent with a broad Atlantic-world confound rather than a slavery-specific bond. More importantly, the relationship disappears once attention is restricted to the 178 dyads with positive mapped exposure (170 in the regression): within the historically exposed set, slave intensity predicts no additional connectedness ($\beta \approx 0.000$,

$p = 1.00$) while ancestry remains positive and marginally significant ($\beta = 0.158$, $p = 0.085$). Heavier trafficking across already-exposed corridors did not generate stronger present-day homeland ties.

The intra-American results are weaker still: in an Americas-only sample of 153 mapped country pairs (151 in the regression), intra-American slave-corridor intensity is negative and insignificant both without ancestry ($\beta = -0.047$, $p = 0.56$) and with it ($\beta = -0.041$, $p = 0.62$). Since this database captures the secondary redistribution of enslaved people within the Americas rather than the Middle Passage, the null appears for subsequent forced removals as well as the original crossing.

A coefficient comparison (Appendix U) places these Atlantic estimates alongside the South African results: voluntary migration produces the strongest homeland-oriented connectedness, indenture an intermediate association, and the forced-migration estimates cluster around zero in both settings. Slavery’s long-run consequences for development and institutions did not take the form of durable positive origin-country social ties in the SCI, and the Atlantic evidence makes the slavery null less dependent on a single national case.

9 Robustness

Inference and permutation. All country-level specifications cluster standard errors on the two country identifiers comprising each undirected dyad (Egger and Tarlea, 2015; Pfaffermayr, 2023); this roughly doubles the baseline ancestry standard error (from 0.017 to 0.031) without changing any key result. County regressions cluster at the state level, and for the small-sample New Zealand and Cape Colony tests I report permutation p -values alongside asymptotic inference. A permutation that shuffles ancestry across pairs places the baseline coefficient 11.1 standard deviations above the null; a network-preserving (node-label) permutation that relabels countries and reassigns the ancestry matrix together leaves it 10.8 standard deviations above, because the two-way country fixed effects already absorb the node-level dependence the node permutation is designed to preserve (Appendix S).

Functional form. The country-level magnitude is sensitive to estimator choice, and I do not treat this as a minor detail. A Park-type test (slope 1.93) indicates Gamma-type rather than Poisson-type heteroskedasticity (Larch, Shikher and Yotov, 2025); a Gamma pseudo-maximum-likelihood estimator yields $\beta = 0.186$ ($p = 0.042$), close to the log-OLS 0.144, whereas standard PPML with country fixed effects yields an insignificant 0.046, and RESET rejects the log-OLS form but not PPML. Because the horse race, permutation test and migrant-stock control all sit inside the same log-linear country-level framework, they share this functional-form sensitivity; the regime contrasts in Sections 6–8, identified from within-country variation in the *type* of migration, do not – which is why the paper places its interpretive weight there. Appendices A and K report the full PPML, Park, Gamma and mean-variance diagnostics.

Competing channels, outcomes and spatial trends. The ancestry coefficient is stable across alternative SCI transformations (asinh, $\log(1 + \text{SCI})$, percentile rank), under winsorisation, and after dropping the most influential pairs. Adding bilateral direct-flight counts (OpenFlights) or restrict-

ing to country pairs above 25% and 50% Facebook penetration leaves it positive and significant (Appendix L), and all kitchen-sink regressors have variance inflation factors below the conventional threshold (Appendix A). Replacing F_{ST} with the coarse Out-of-Africa migratory-distance proxy of Ashraf and Galor (2013) leaves the coefficient essentially unchanged, and the result is robust in sign across alternative ancestry constructions, with the log transformation the most precisely estimated. Finally, the subnational designs are largely insulated from the spatial-trend critique of Conley and Kelly (2025) and the pitfalls catalogued by Arroyo Abad and Maurer (2021): the county, New Zealand and South African exercises identify from within-state or within-country variation rather than cross-country gradients, and use dated historical sources rather than compressing centuries into a single treatment. The full versions of these checks are in Appendix S.

10 Discussion and Conclusion

The horse race establishes that shared ancestry captures something beyond genetic distance, linguistic proximity, religious similarity, cultural values, folklore traditions and historical trade. What remains? My interpretation is that the residual ancestry channel is consistent with what I call, descriptively, the *social infrastructure* of migration: the diaspora networks, family ties, institutional transplants and cultural memory that can persist across generations after mass population movements. When British settlers moved to New Zealand, they carried not only language and institutions but also webs of family connection, association and memory that linked colony and metropole. Those links may be reproduced across generations through families, churches, ethnic organisations, alumni networks and cultural practices. I use *social infrastructure* as an organising frame for these patterns, not as a separately identified mechanism: the analysis is descriptive throughout, and the term should be read as an interpretation of the residual association.

This clarifies the paper’s place in the cultural-transmission literature: migration can transmit not only preferences or identities within a community but durable cross-border social ties, a pattern consistent with work emphasising endogenous socialization and local institutions (Bisin and Verdier, 2023; Giuliano and Nunn, 2021). The US and New Zealand exercises are useful because they show this below the country level, where transmission mechanisms are easier to describe; and it is not confined to the European settler context – Ma and Fu (2025) find that origin-destination genetic distance within China reduces migrants’ settlement intention.

This interpretation does not claim that ancestry operates independently of material communication infrastructure. Maintaining homeland ties has always depended on transport, literacy, postal and shipping routes, telecommunications, wealth and legal mobility, and several controls speak to these imperfectly: historical trade proxies shipping intensity, migrant stocks current diaspora links, direct flights modern travel, the country fixed effects every origin- and destination-specific characteristic including GDP and population, and the South African regressions add internet, dwelling and urbanisation controls. These do not fully solve the problem – such channels are likely mediators as much as confounders, since voluntary settler communities often built precisely the infrastructure

through which ancestry became durable. The type-of-migration evidence matters for exactly this reason: slavery and other coerced movements disrupted families, institutions, resources and communication at once, so the empirical contrast identifies the long-run difference between migration regimes, not a single isolated component.

Several caveats apply, and Appendix T discusses them in full. The analysis is descriptive rather than causal; the Putterman matrix is an estimate, not a census, and is coarsest precisely for forced-migration origins; ACS ancestry is self-reported; and the SCI captures friendship links on the world’s largest open social network only, leaving closed platforms such as WeChat and VK – and hence China and Russia – unobserved, and offering a single cross-section rather than a time series.

Five centuries of human migration have left a durable imprint on 21st-century digital social networks. Shared ancestry predicts bilateral Facebook friendships across countries – surviving a horse race against genetics, language, religion, folklore, trade and contemporary migrant stocks – and three subnational analyses show that the relationship holds within countries and operates at fine geographic scales. The strongest evidence on the *conditions* of migration comes from the South African regime contrast and the US county patterns; the Australasian convict comparison points in the same direction but is imprecisely estimated in a five-colony sample, and New Zealand’s strictest partner-specific test does not survive demanding fixed effects. The evidence is descriptive throughout, and the country-level magnitude is sensitive to estimator choice, but the central pattern across the better-powered designs is consistent: voluntary migration builds lasting cross-border social ties, and coerced migration does not.

Each of the paper’s four contributions closes a specific gap. For the economics of slavery and coercion, the paper identifies what forced migration failed to create. The South African comparison shows that voluntary settlement, indenture and slavery generated sharply different patterns of modern homeland connectedness; the Atlantic and intra-American extensions confirm that the slavery null is not confined to one country; and the Australian convict comparison gives directionally consistent point estimates without rejecting the null in a five-colony sample, leaving the cross-country case for Hypothesis 3 resting primarily on the South African and Atlantic evidence. For the migration literature, the paper adds a bilateral relational legacy to the set of outcomes through which migration reshapes economies: ancestry predicts the geography of everyday social connectedness, the substrate on which investment, tourism and information flows can travel. For the cultural transmission literature, the paper identifies a concrete object of intergenerational transmission – what I call social infrastructure – and shows that its persistence varies with the agency available to migrants, as the Bisin-Verdier framework would predict. For the deep-roots and persistence literatures, the paper offers a new outcome – network ties rather than income or institutions – and addresses the spatial-trend critique of Conley and Kelly (2025) by identifying the relationship through within-state and within-country designs rather than cross-country gradients.

Several extensions would sharpen the results. First, the SCI is available at GADM2 district-to-district resolution globally. Exploiting that finer geography – particularly for New Zealand and

South Africa, where the current tests use district-to-country aggregates – could tighten the within-country estimates and allow richer partner-specific tests. Second, the Indian indentured records are only partially matched to districts; a fuller district-by-origin implementation would strengthen the intermediate case that currently rests on modern population shares. Third, other forced-migration settings offer natural out-of-sample tests. US slavery produced massive internal forced movement whose demographic traces are visible at the county level; Caribbean plantation economies received enslaved populations from specific African regions documented in the Slave Voyages data. Whether those settings replicate the Cape Colony null at subnational scale remains an open question. Fourth, if Meta releases updated SCI data, repeated cross-sections could eventually reveal whether ancestry-connectedness is stable or eroding as the migration events recede further into history.

The broader point is that not all migration is equal. The Putterman-Weil matrix treats five centuries of population flows as a single demographic accounting, but the evidence here suggests that the conditions of migration – voluntary or coerced, concentrated or dispersed, institutionally embedded or institutionally excluded – determine whether demographic history becomes a durable social bridge. If the current wave of global migration follows the same logic, it will leave social traces that persist long after the migration events themselves have receded into history. But the Cape Colony reminds us that this future depends on the conditions of migration. The social legacies of voluntary movement and forced displacement are not the same.

Tables (continued)

Table 4: Historical birthplace shares predict 2021 Facebook SCI: New Zealand

	Panel A: NZ districts (GADM2)				Panel B: NZ stacked district $\times$ partner			
	(D1) %UK $\rightarrow$ GB	(D2) %Scot $\rightarrow$ GB	(D3) %Irish $\rightarrow$ IE	(D4) %UK $\rightarrow$ GB + FE	(S1) Matched	(S2) Matched + PF	(S3) Matched + Other + PF	(S4) Matched + DF + PF
<i>Dep. var.: log(SCI)</i>								
% UK-born	0.317*** (0.082)			-0.896 (1.018)				
% Scottish-born		0.199 (0.119)						
% Irish-born			0.173 (0.160)					
Matched birthplace share					0.225*** (0.031)	-0.031 (0.023)	-0.033 (0.023)	-0.001 (0.012)
Other tracked origins							0.144. (0.076)	
log(pop)	-0.081 (0.078)	-0.092 (0.081)	-0.060 (0.097)	-0.113 (0.093)	-0.093 (0.097)	-0.045 (0.029)	-0.040 (0.028)	
Region FE	No	No	No	Yes	No	No	No	No
Partner FE	No	No	No	No	No	Yes	Yes	Yes
District FE	No	No	No	No	No	No	No	Yes
R^2	0.098	0.133	0.039		0.172	0.930	0.931	0.975
Within R^2				0.187		0.030	0.043	0.000
N	29	29	29	23	203	203	203	203
Perm. p	0.143							0.944

Notes: Dependent variable: log(SCI). Panel A reports NZ-only district regressions using 29 GADM2 districts built from 65 borough-level birthplace tabulations in the 1881 Census. Column (D4) adds GADM1 region fixed effects and is therefore identified only from within-region variation. Panel B stacks the same 29 districts against seven tracked partner countries. “Matched birthplace share” is the historical share born in the partner country for that district-partner observation. “Other tracked origins” is the sum of the remaining tracked birthplace shares. PF = partner fixed effects; DF = district fixed effects. Perm. p : two-sided permutation p -value from 5,000 permutations. Heteroskedasticity-robust standard errors. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.1$.

Table 5: South Africa: population composition predicts social connectedness (52 districts)

	(S1) Baseline	(S2) +Lang	(S3) Kitchen	(S4) India	(S5) Af. plac.	(S6) Slave orig.
<i>Dep. var.: log(SCI)</i>						
% White	0.578*** (0.014)	0.548*** (0.015)	0.404*** (0.017)		0.230*** (0.030)	
Common lang.		0.479*** (0.062)	0.372*** (0.064)			
% Coloured			0.125*** (0.012)			0.158*** (0.016)
% Indian			0.040** (0.013)	0.078** (0.025)		
log(dist)				-1.117 (0.616)	-0.723*** (0.090)	-0.878*** (0.131)
Partner FE	Yes	Yes	Yes	No	Yes	Yes
Within R^2	0.705	0.724	0.766	—	0.332	0.241
N	728	728	728	52	312	364

Notes: Dependent variable: $\log(SCI)$. The population and population-composition regressors (%White, %Coloured, %Indian) enter as logged shares, $\log(\%)$, so the reported coefficients are elasticity-like associations rather than level-share effects. All specifications include $\log(\text{pop})$. Distance is omitted from the European specifications (S1–S3) because all SA districts are approximately equidistant from European capitals; $\log(\text{dist})$ is included in the India (S4), African-neighbour (S5) and slave-origin (S6) specifications, where district-to-partner distance varies meaningfully. EU partners: GB, NL, DE, FR, PT, BE, IT, ES, CH, AT, IE, SE, NO, DK. African neighbours: MZ, LS, SZ, ZM, TZ, MW. Slave-origin countries: MZ, MG, ID, IN, MY, LK, MU. Heteroskedasticity-robust standard errors. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table 6: Cape Colony: historical settlement origins predict social connectedness (14 districts)

	(C1) Baseline	(C2) +1820	(C3) +Hug	(C4) +Both	(C5) Slave	(C6) Kitchen
<i>Dep. var.: log(SCI)</i>						
% White	0.785*** (0.027)	0.698*** (0.032)	0.721*** (0.033)	0.680*** (0.029)		0.562*** (0.029)
1820-settler surnames		0.042*** (0.005)		0.098*** (0.020)		0.088*** (0.020)
Huguenot surnames			0.029*** (0.005)	−0.045** (0.017)		−0.046** (0.017)
Slave origins					0.010 (0.008)	−0.030**† (0.010)
Partner FE	Yes	Yes	Yes	Yes	Yes	Yes
Within R^2	0.789	0.839	0.830	0.844	0.190	0.722
N	196	196	196	196	98	294

Notes: Dependent variable: $\log(SCI)$. The %White, %Coloured and owner-surname (1820-settler, Huguenot) and slave-origin regressors all enter as logged shares, $\log(\% \cdot)$, so the coefficients are elasticity-like associations. All specifications include $\log(\text{pop})$ and common language. C5 (slave-origin baseline, 98 obs) includes $\log(\text{dist})$. C6 pools all 21 European and slave-origin partners (294 obs) and, like the European columns, omits $\log(\text{dist})$. †The slave-origin coefficient in C6 reflects a single pooled $\log(\% \text{White})$ slope across both partner groups; once the European proxies are interacted with an EU-region dummy it is -0.003 (n.s.), and re-estimated on the slave-origin observations only (with $\log(\text{dist})$) it is -0.012 (SE 0.007, $p = 0.11$). The strong negative is therefore a pooling artifact; see Section 8.3. 1820-settler and Huguenot owner-surname shares constructed from the slave emancipation records of Ekama et al. (2021), aggregated from 11 colonial-Cape districts to eight modern GADM2 districts using a manual crosswalk; remaining Cape Colony districts enter with a small floor share. Slave origins from the same emancipation dataset, matched to the farm boundaries of Fourie (2022) (914 records). Farm-name robustness specifications are reported in Appendix N. Heteroskedasticity-robust standard errors. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.1$.

Figures

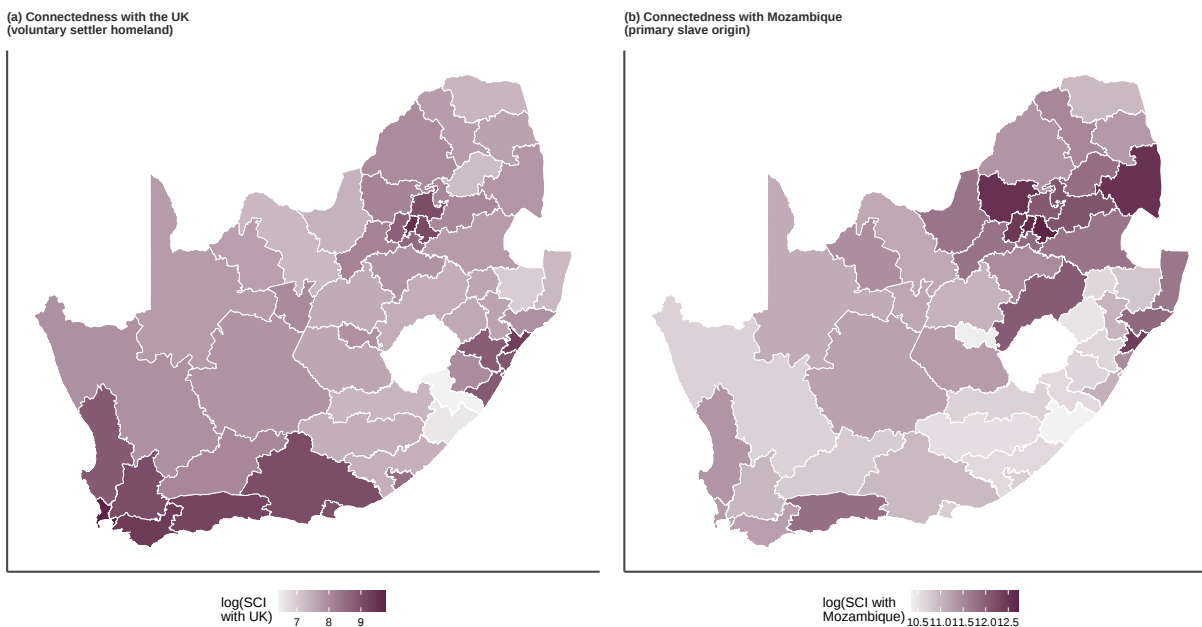


Figure 2: Voluntary versus forced migration in South Africa. Panel (a) shows Facebook connectedness with the United Kingdom – the primary homeland of voluntary settlers – by GADM2 district. Panel (b) shows connectedness with Mozambique – the dominant origin of enslaved people imported to the Cape Colony. The strong spatial gradient in panel (a) reflects the geographic distribution of European-descended populations: districts with more White residents (Western Cape, Gauteng) are substantially more connected to the UK. The absence of a comparable gradient in panel (b) is consistent with Hypothesis 3: forced migration through slavery did not create lasting social connections with origin countries.

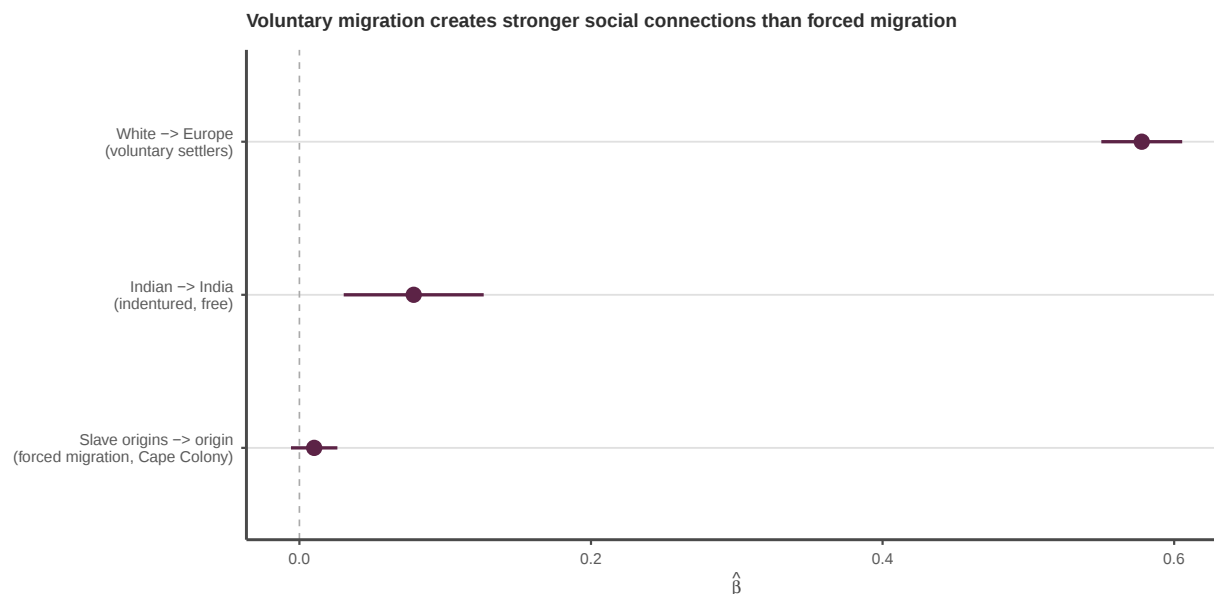


Figure 3: Free versus forced migration: coefficient comparison. Voluntary settler ancestry (White → Europe) creates strong connections; indentured migration (Indian → India) a moderate one; forced migration through slavery (slave origins → origin countries) no comparable positive relationship.

Data availability

The full analysis code and a documented, linked list of every data source are available at https://github.com/johanfourieza/research/tree/main/2026_uprooted. The primary data are publicly available third-party datasets, including Meta’s Social Connectedness Index (via the Humanitarian Data Exchange), the Putterman and Weil World Migration Matrix, the CEPII gravity, distance and trade databases, the Slave Voyages database, the American Community Survey, GADM administrative boundaries, and national census sources for New Zealand, Australia and South Africa. Two restricted datasets – the Cape Colony farm boundaries and the Natal Indian indentured records – are available from the author on request.

References

- Abramitzky, Ran, and Leah Boustan.** 2022. *Streets of Gold: America’s Untold Story of Immigrant Success*. PublicAffairs.
- Acemoglu, Daron, and Alexander Wolitzky.** 2011. “The Economics of Labor Coercion.” *Econometrica*, 79(2): 555–600.
- Anderson, Clare, and Hamish Maxwell-Stewart.** 2014. “Convict Labour and the Western Empires, 1415–1954.” In *The Routledge History of Western Empires*, ed. Robert Aldrich and Kirsten McKenzie, 102–117. London:Routledge.

- Aneja, Abhay, and Guo Xu.** 2024. “Strengthening State Capacity: Civil Service Reform and Public Sector Performance during the Gilded Age.” *American Economic Review*, 114(8): 2352–2387.
- Arroyo Abad, Leticia, and Noel Maurer.** 2021. “History Never Really Says Goodbye: A Critical Review of the Persistence Literature.” *Journal of Historical Political Economy*, 1: 31–68.
- Ashraf, Quamrul, and Oded Galor.** 2013. “The “Out of Africa” Hypothesis, Human Genetic Diversity, and Comparative Economic Development.” *American Economic Review*, 103(1): 1–46.
- Australian Bureau of Statistics, Australian Data Archive, Len Smith, Tim Rowse, and Stuart Hungerford.** 2019. “Historical and Colonial Census Data Archive (HCCDA).” Digitised census tables from Australian colonies, 1833–1901.
- Bailey, Michael, Abhinav Gupta, Sebastian Hillenbrand, Theresa Kuchler, Robert Richmond, and Johannes Stroebel.** 2021. “International Trade and Social Connectedness.” *Journal of International Economics*, 129: 103418.
- Bailey, Michael, Drew M. Johnston, Theresa Kuchler, Johannes Stroebel, and Arlene Wong.** 2022. “Peer Effects in Product Adoption.” *American Economic Journal: Applied Economics*, 14(3): 488–526.
- Bailey, Michael, Ruiqing Cao, Theresa Kuchler, Johannes Stroebel, and Arlene Wong.** 2018. “Social Connectedness: Measurement, Determinants, and Effects.” *Journal of Economic Perspectives*, 32(3): 259–280.
- Ballantyne, Tony.** 2014. “The Theory and Practice of Empire-Building: Edward Gibbon Wakefield and ‘Systematic Colonisation’.” In *The Routledge History of Western Empires.*, ed. Robert Aldrich and Kirsten McKenzie, 89–101. London: Routledge.
- Bandiera, Oriana, Imran Rasul, and Martina Viarengo.** 2013. “The Making of Modern America: Migratory Flows in the Age of Mass Migration.” *Journal of Development Economics*, 102: 23–47.
- Bazzi, Samuel, Martin Fiszbein, and Mesay Gebresilasie.** 2020. “Frontier Culture: The Roots and Persistence of “Rugged Individualism” in the United States.” *Econometrica*, 88(6): 2329–2368.
- Becker, Sascha O., Katrin Boeckh, Christa Hainz, and Ludger Woessmann.** 2016. “The Empire Is Dead, Long Live the Empire! Long-Run Persistence of Trust and Corruption in the Bureaucracy.” *Economic Journal*, 126(590): 40–74.
- Bisin, Alberto, and Thierry Verdier.** 2001. “The Economics of Cultural Transmission and the Dynamics of Preferences.” *Journal of Economic Theory*, 97(2): 298–319.

- Bisin, Alberto, and Thierry Verdier.** 2023. “Advances in the Economic Theory of Cultural Transmission.” *Annual Review of Economics*, 15: 63–89.
- Blumenstock, Joshua E., Guanghua Chi, and Xu Tan.** 2025. “Migration and the Value of Social Networks.” *Review of Economic Studies*, 92(1): 97–128.
- Burchardi, Konrad B., Thomas Chaney, and Tarek A. Hassan.** 2019. “Migrants, Ancestors, and Foreign Investments.” *Review of Economic Studies*, 86(4): 1448–1486.
- Cameron, A. Colin, and Douglas L. Miller.** 2015. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources*, 50(2): 317–372.
- Cavalli-Sforza, Luigi Luca, and Marcus W. Feldman.** 1981. *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton University Press.
- Conley, Timothy G., and Morgan Kelly.** 2025. “The Standard Errors of Persistence.” *Journal of International Economics*, 153: 104027.
- Cowley, Trudy, Lucy Frost, Kris Inwood, Rebecca Kippen, Hamish Maxwell-Stewart, Monika Schwarz, John Shepherd, Richard Tuffin, Mark Williams, John Wilson, and Paul Wilson.** 2023. “Reconstructing a Longitudinal Dataset for Tasmania.” In *Sowing: The Construction of Historical Longitudinal Population Databases.*, ed. Kees Mandemakers, George Alter, Hélène Vézina and Paul Puschmann, 455–482. Nijmegen:Radboud University Press.
- Desmet, Klaus, and Romain Wacziarg.** 2021. “The Cultural Divide.” *Economic Journal*, 131(637): 2058–2088.
- Donaldson, Dave, and Richard Hornbeck.** 2016. “Railroads and American Economic Growth: A “Market Access” Approach.” *Quarterly Journal of Economics*, 131(2): 799–858.
- Edmonds, Penelope, and Jane Carey.** 2017. “Australian Settler Colonialism over the Long Nineteenth Century.” In *The Routledge Handbook of the History of Settler Colonialism.*, ed. Edward Cavanagh and Lorenzo Veracini, 371–390. Abingdon:Routledge.
- Egger, Peter H., and Filip Tarlea.** 2015. “Multi-way clustering estimation of standard errors in gravity models.” *Economics Letters*, 134: 144–147.
- Ekama, Kate, Johan Fourie, Hans Heese, and Lisa-Cheree Martin.** 2021. “When Cape Slavery Ended: Introducing a New Slave Emancipation Dataset.” *Explorations in Economic History*, 81: 101390.
- Fouquin, Michel, and Jules Hugot.** 2016. “Two Centuries of Bilateral Trade and Gravity Data: 1827–2014.” CEPII CEPII Working Paper No. 2016-14.

- Fourie, Johan.** 2022. “The Settlers of South Africa: Economic Forces of the Expanding Frontier.” In *Migration in Africa: Shifting Patterns of Mobility from the 19th to the 21st Century.*, ed. Michiel de Haas and Ewout Frankema, 135–154. Routledge.
- Fourie, Johan, and Maria Santana-Gallego.** 2013. “Ethnic Reunion and Cultural Affinity.” *Tourism Management*, 36: 411–420.
- Giuliano, Paola, and Nathan Nunn.** 2021. “Understanding Cultural Persistence and Change.” *Review of Economic Studies*, 88(4): 1541–1581.
- Head, Keith, Thierry Mayer, and John Ries.** 2010. “The Erosion of Colonial Trade Linkages after Independence.” *Journal of International Economics*, 81(1): 1–14.
- Larch, Mario, Serge Shikher, and Yoto V. Yotov.** 2025. “Estimating Gravity Equations: Theory Implications, Econometric Developments, and Practical Recommendations.” *Review of International Economics*, 33(5): 1066–1092.
- Ma, Junfeng, and Wentao Fu.** 2025. “Ancestors Call Me Home: Cultural Distance and Settlement Intention of Domestic Migrant Workers in China.” *Cities*, 167: 106317.
- Michalopoulos, Stelios, and Melanie Meng Xue.** 2021. “Folklore.” *Quarterly Journal of Economics*, 136(4): 1993–2046.
- Nicholas, Stephen,** ed. 1988. *Convict Workers: Reinterpreting Australia’s Past.* Cambridge:Cambridge University Press.
- Nunn, Nathan.** 2008. “The Long-Term Effects of Africa’s Slave Trades.” *Quarterly Journal of Economics*, 123(1): 139–176.
- Nunn, Nathan, and Leonard Wantchekon.** 2011. “The Slave Trade and the Origins of Mistrust in Africa.” *American Economic Review*, 101(7): 3221–3252.
- Obradovich, Nick, Omer Ozak, Ignacio Martin, Ignacio Ortuno-Ortin, Edmond Awad, Manuel Cebrian, Ruben Cuevas, Klaus Desmet, Iyad Rahwan, and Angel Cuevas.** 2022. “Expanding the Measurement of Culture with a Sample of Two Billion Humans.” *Journal of the Royal Society Interface*, 19: 20220085.
- Olssen, Erik.** 2025. *The Origins of an Experimental Society: New Zealand, 1769–1860.* Auckland:Auckland University Press.
- Oster, Emily.** 2019. “Unobservable Selection and Coefficient Stability: Theory and Evidence.” *Journal of Business & Economic Statistics*, 37(2): 187–204.
- Pfaffermayr, Michael.** 2023. “Cross-sectional Gravity Models, PPML Estimation, and the Bias Correction of the Two-Way Cluster-Robust Standard Errors.” *Oxford Bulletin of Economics and Statistics*, 85(5): 1111–1134.

- Putterman, Louis, and David N. Weil.** 2010. “Post-1500 Population Flows and the Long-Run Determinants of Economic Growth and Inequality.” *Quarterly Journal of Economics*, 125(4): 1627–1682.
- Santos Silva, João M. C., and Silvana Tenreyro.** 2006. “The Log of Gravity.” *Review of Economics and Statistics*, 88(4): 641–658.
- Sequeira, Sandra, Nathan Nunn, and Nancy Qian.** 2020. “Immigrants and the Making of America.” *Review of Economic Studies*, 87(1): 382–419.
- SlaveVoyages Consortium.** 2024. “Slave Voyages.” <https://www.slavevoyages.org/>, Accessed data from the trans-Atlantic and intra-American slave trade databases.
- Spolaore, Enrico, and Romain Wacziarg.** 2009. “The Diffusion of Development.” *Quarterly Journal of Economics*, 124(2): 469–529.
- Spolaore, Enrico, and Romain Wacziarg.** 2013. “How Deep Are the Roots of Economic Development?” *Journal of Economic Literature*, 51(2): 325–369.
- Spolaore, Enrico, and Romain Wacziarg.** 2016. “Ancestry, Language and Culture.” In *The Palgrave Handbook of Economics and Language.*, ed. Victor Ginsburgh and Shlomo Weber, 174–211. Hampshire:Palgrave Macmillan.
- Spolaore, Enrico, and Romain Wacziarg.** 2018. “Ancestry and Development: New Evidence.” *Journal of Applied Econometrics*, 33(5): 748–762.
- Statistics South Africa.** 2024. “South African Census 2022, 10% sample [dataset]. Version 1.” *Pretoria: Statistics South Africa [producer]; Cape Town: DataFirst [distributor]*.
- Tabellini, Marco.** 2020. “Gifts of the Immigrants, Woes of the Natives: Lessons from the Age of Mass Migration.” *Review of Economic Studies*, 87(1): 454–486.
- United Nations Department of Economic and Social Affairs.** 2024. “International Migrant Stock 2024.” Bilateral migrant stock matrix, mid-year 2020 estimates.
- US Census Bureau.** 2022. “American Community Survey 2018–2022 5-Year Estimates, Table B04006: People Reporting Ancestry.” <https://data.census.gov>.
- Valencia Caicedo, Felipe.** 2019. “The Mission: Human Capital Transmission, Economic Persistence, and Culture in South America.” *Quarterly Journal of Economics*, 134(1): 507–556.
- Waetjen, Thembisa, and Goolam Vahed.** 2014. “Passages of Ink: Decoding the Natal Indentured Records into the Digital Age.” *Kronos*, 40(1): 45–73.
- Whatley, Warren.** 2022. “How the International Slave Trades Underdeveloped Africa.” *Journal of Economic History*, 82(2): 403–441.

Wright, Gavin. 2022. “Slavery and the Rise of the Nineteenth-Century American Economy.”
Journal of Economic Perspectives, 36(2): 123–148.