Online Appendix for: Place, Peers, and the Teenage Years

By Nathan Deutscher

Appendix A: Additional charts

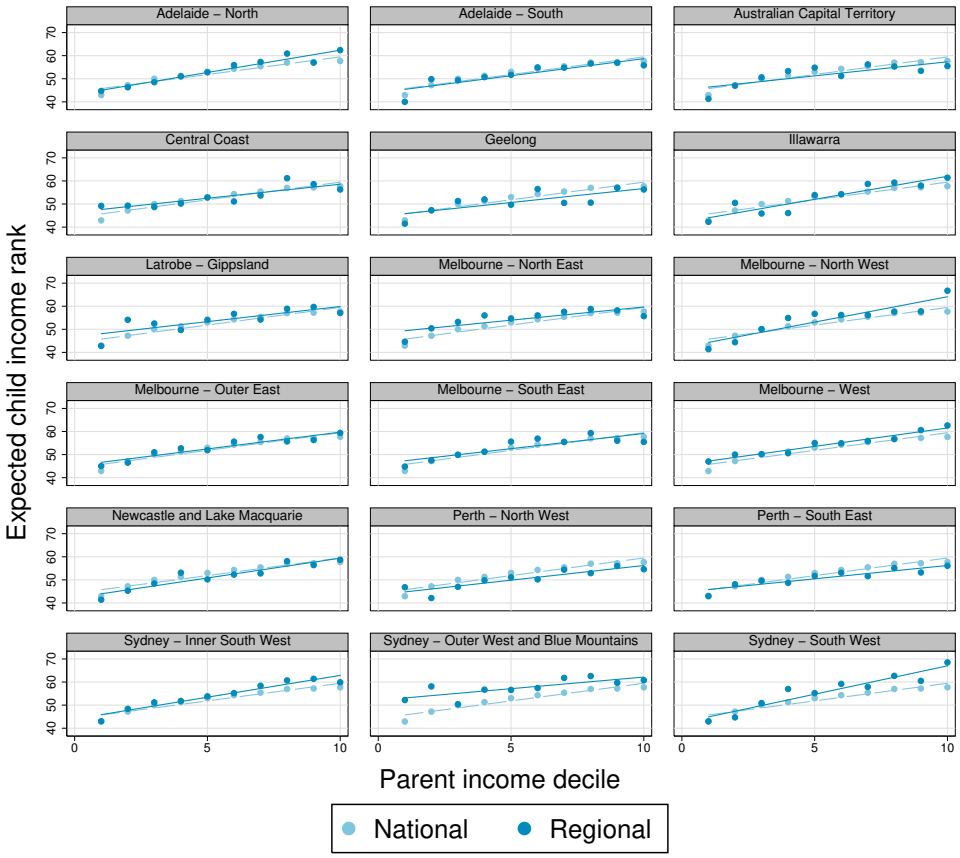


Figure A.1. National and regional relationships between parent and child income ranks: permanent residents born in 1978

Notes: Based on the sample of permanent residents. Chart illustrates the mean household total income rank at age 24, by parent income decile, for children born in 1978 and in one of the 18 largest SA4.

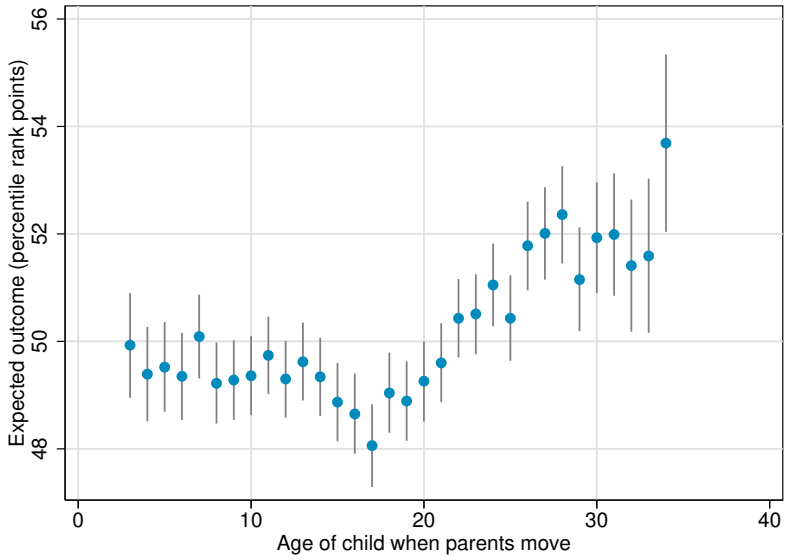


FIGURE A.2. EXPECTED OUTCOME OF CHILD: BORN IN 1991, TO PARENTS WITH MEDIAN INCOME, AND MOVING BETWEEN PLACES WHERE SIMILAR PERMANENT RESIDENTS END UP WITH MEDIAN INCOME

Notes: Estimated linear combination $\alpha^1_{1991} + 50\alpha^2_{1991} + \zeta^1_m + 50\zeta^2_m$ of coefficients from equation (3). This captures the expected household income rank at age 24 for a child: born in 1991; with parents at the 50^{th} percentile of the income distribution; and moving at age m between an origin and destination where their predicted outcome based on permanent residents is also the 50^{th} percentile. The full regression regresses the adult ranks y_i of those whose parents move once in their childhood on the interaction of their age at parent move m with $\Delta_{odps} = \bar{y}_{dps} - \bar{y}_{ops}$ — the difference between the expected outcomes for permanent residents of the same parent percentile rank p and cohort s in the destination d versus the origin o. Controls capture: cohort and origin effects (via indicators for cohort and their interactions with predicted outcomes for permanent residents of the origin); disruption effects (via indicators for age at move and their interaction with parental rank); and indicators for cohort interacted with Δ_{odps} to capture potential mis-measurement of the origin.

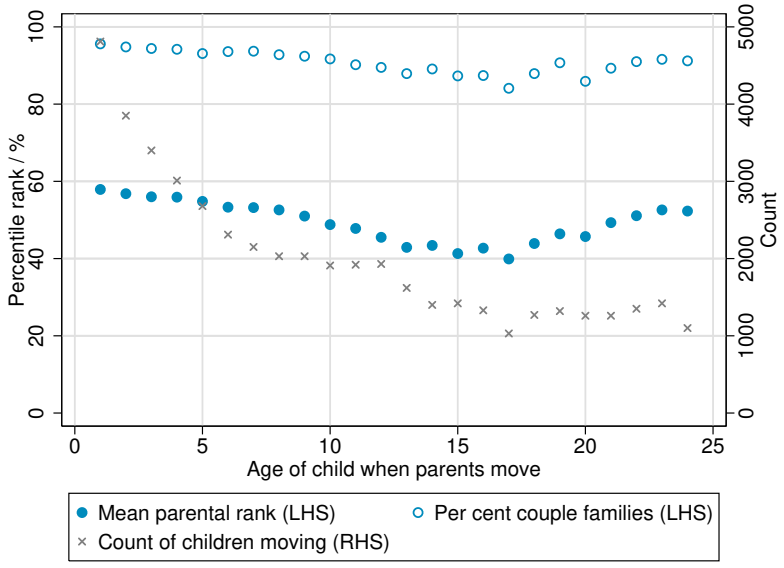


FIGURE A.3. FAMILY CHARACTERISTICS BY AGE AT MOVE: 1991 COHORT

Notes: For the individuals born in the 1991 financial year whose parents move once, shows the mean parent rank, proportion in couple families and sample size by the individual's age at move.

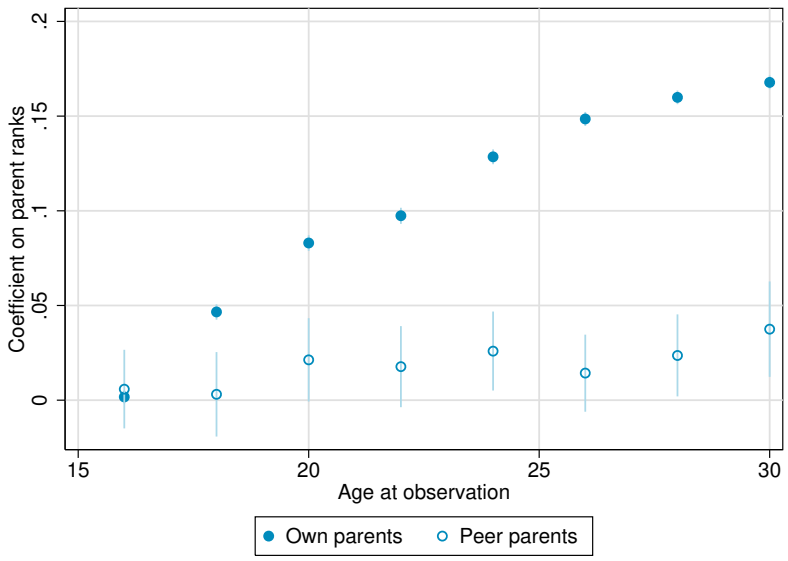


FIGURE A.4. INFLUENCE OF OWN AND PEER PARENTS ON HOUSEHOLD INCOME RANK AT VARIOUS AGES

Notes: Based on permanent postcode residents. Shows the coefficients (and 95% confidence intervals) from a regression of household income rank at various ages on own parent household income rank and the mean parent household income rank of peers (defined by shared permanent postcode and financial year of birth). A 7-year moving average of the mean parent rank of peers is included as a control, in line with the specification in column (1) of Table 4.

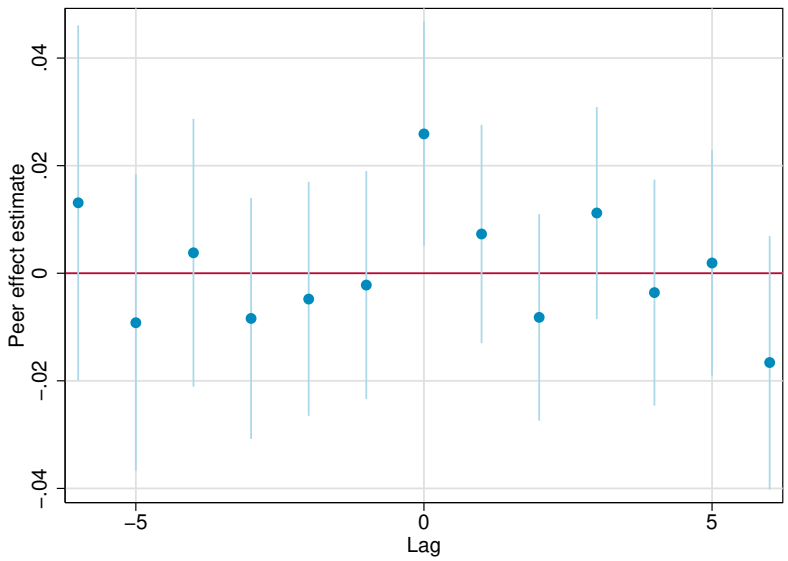


FIGURE A.5. PEER EFFECT ESTIMATES: PLACEBO TEST

Notes: Based on permanent postcode residents. Shows the coefficients (and 95% confidence intervals) from a regression of household income rank at age 24 on own parent household income rank and the mean parent household income rank of peers (defined by shared permanent postcode and a financial year of birth that is shifted by a lag l relative to the individual's own). A 7-year moving average of the mean parent rank of peers is included as a control, in line with the specification in column (1) of Table 4. The lack of an effect for surrounding years may be a little surprising given some peers will likely be drawn from adjacent birth cohorts. However, the design is not well suited to identifying the effect of temporally adjacent peer groups, as it relies on these peer groups to identify idiosyncratic variation. Thus a relatively high income peer group for the postcode in one year will be correlated with relatively lower income peer groups in the years either side. When examining the effect of a birth cohort that is not your own, the effect of a richer cohort will conceivably be masked by the effect of a poorer own cohort, from which the majority of your peers are actually drawn. To put it another way, the downward bias that may apply to the moving average approach as discussed in Black, Devereux and Salvanes (2013) will be magnified when looking at a birth cohort that is not your own. This exercise is best seen as a test of the specification rather than an attempt to credibly identify the causal effect of the peers born in the years either side.

Appendix B: Additional tables

Table B.1—Summary statistics for permanent residents and one-time movers

	Permanent residents			1-time movers		
	Mean	Std. dev.	Median	Mean	Std. dev.	Median
Panel A: Family background						
Parent income ($)	79,300	71,600	72,600	86,100	80,700	77,700
Parent income rank	50.7	28.5	51	54.8	27.3	56
Indicator, in a couple family	0.87	0.33	1	0.91	0.29	1
Family size	2.7	1.2	3	2.7	1.2	2
Panel B: Outcomes						
Child income ($)	61,800	46,400	54,100	62,600	45,300	54,800
Child rank	52.2	28.4	53	52.8	28.6	54
N	1,683,800			313,900		

Notes: The full sample consists of those children born between 1978-91, remaining resident in Australia through to 2015 and linked to parents. The permanent residents are those children whose primary parent files from only one SA4 from 1991 through to the year the child turned 35. The 1-time movers are those whose primary parent filed from two SA4 from 1991 through to the year the child turned 35, filed from each at least twice, began filing in the destination the year after they ceased filing in the origin, and moved at least 15 kilometres (based on postcode centroids). Parent income is the average household total pre-tax income from 1991-2001 in 2015 dollars. Child income is the household total pre-tax income in the year the child turns 24. Ranks are calculated separately for each birth cohort.

Table B.2—Difference between destination and origin: 1-time mover subsample

	Mean	Std. dev.	Median
Mean permanent resident parent rank	-1.06	9.98	-.83
Number of permanent residents	-95	1,000	-57
Predicted child rank	-.083	5.09	-.088
N	313,900		

Notes: Shows differences in the characteristics of the 1-time movers destinations and origins. These characteristics of place are based on the permanent residents. The difference in the predicted child rank is simply the difference in predicted values for a child in birth cohort s and with parent income rank p for a permanent resident of the origin o versus the destination d, that is $\Delta_{odps} = \bar{y}_{dps} - \bar{y}_{ops}$. The difference in the mean permanent resident parent rank is the difference in the means of the parent ranks p of the permanent residents in the same cohort s in the origin o and destination d, that is $\bar{p}_{os} - \bar{p}_{ds}$.

Table B.3—Exposure effect estimates and model fit statistics: by model specification

	Linear	Piecewise linear with kink at age...						
		10	11	12	13	14	15	16
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	0.033							
	(0.002)							
Early		0.010	0.011	0.015	0.018	0.020	0.023	0.026
		(0.008)	(0.007)	(0.006)	(0.005)	(0.005)	(0.004)	(0.004)
Late		0.039	0.042	0.043	0.043	0.045	0.045	0.045
		(0.003)	(0.003)	(0.004)	(0.004)	(0.005)	(0.005)	(0.006)
Post-outcome	0.008	0.008	0.008	0.008	0.009	0.009	0.009	0.009
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
$(R^2 - R^2_{max})10^6$	-10	-2	0	-1	-2	-2	-4	-6
$(aR^2 - aR^2_{max})10^6$	-9	-2	0	-1	-2	-2	-4	-6
$AIC_{min} - AIC$	-10	-3	0	-1	-3	-3	-5	-7
$BIC_{min} - BIC$	0	-4	-1	-2	-3	-3	-6	-8
N	264,500	264,500	264,500	264,500	264,500	264,500	264,500	264,500

Notes: Exposure effect estimates and model fit statistics for competing models of exposure effects — a constant exposure effects model as in Chetty and Hendren (2018) and a piecewise linear model with the kink at varying ages. Model fit statistics are transformed as described to aid readability — higher values indicate better fits. Statistics are estimated from equation (4) for early ($m \in \{2, ..., k\}$), late ($m \in \{k, ..., 24\}$) or post-outcome ($m \in \{25, ..., 34\}$) exposure for varying values of the kink k (columns (2)-(8)) or assuming constant exposure effects in early and late childhood (column (1)). The coefficients represent the expected boost to an individual's household income rank associated with an additional year at this stage of life in a destination with 1 percentile rank higher expected outcomes for permanent residents. They are estimated by regressing the adult ranks y_i of those whose parents move once in their childhood on the interaction of their time exposed to the destination at each life stage with $\Delta_{odps} = \bar{y}_{dps} - \bar{y}_{ops}$ — the difference between the expected outcomes for permanent residents of the same parent percentile rank p and cohort s in the destination d versus the origin o. Controls capture: cohort and origin effects (via indicators for cohort and their interactions with predicted outcomes for permanent residents of the origin); disruption effects (via indicators for age at move and their interaction with parental rank); and indicators for cohort interacted with Δ_{odps} to capture potential mis-measurement of the origin. Murphy-Topel standard errors are in parentheses.

Table B.4—Parent and peer influences on household income rank at age 24: robustness of moving average specification

	(1)	(2)	(3)	(4)	(5)	(6)
Parent rank	0.131	0.131	0.130			
	(0.002)	(0.002)	(0.002)			
Peers	0.017	0.026	0.030	0.020	0.028	0.032
	(0.008)	(0.010)	(0.011)	(0.020)	(0.023)	(0.027)
Specification						
Window width	3	5	7	3	5	7
Postcode controls	X	X	X			
Family fixed effects				X	X	X
N	1,040,900	854,700	670,000	1,126,200	939,700	754,900

Notes: Coefficients from equation (5) — the regression of a child's household income rank at age 24 on: their parent household income rank; and their peers mean parent rank; the 3-, 5- or 7-year moving average of the same; and additional controls. These additional controls are either: postcode and cohort fixed effects, a postcode linear trend and the postcode's mean government benefits paid, higher education loan debt, salary and wages, and total income for each individual with a tax liability in the year of observation (1)-(3); or family fixed effects (4)-(6). Peers are defined by postcode and financial year of birth and exclude the individual in question. A peer's primary parent must have been a permanent resident of the postcode — not filing from outside it — from 1991 to the year in which the child turned 20. Robust standard errors, clustered by postcode, are in parentheses.

Appendix C: Validation exercises

This Appendix replicates validation exercises conducted by Chetty and Hendren (2018), with largely comforting results. The first set of tests considers the robustness of the estimates to more general specifications and later ages of observation, the remainder examine in more detail the key identifying assumption — that selection effects do not vary with the age at move of the child.

C1. Specification and age at observation

In Figure C.1 I show that the patterns of exposure effects observed in Figure 1 emerge even if using the more general specification in equation (2). This more general specification replaces parametric controls for origin and disruption effects with fixed effects for each combination of parent income decile, cohort, origin and age at move. Age-invariant selection effects, positive exposure effects and the pronounced sensitivity of the teenage years all remain apparent.

In Table C.2 I switch attention to the models in which place effects are explicitly modeled as a function of exposure to place. Once again moving from the baseline model (column (1)) to one where parametric controls for origin and disruption effects are replaced by fixed effects (column (2)) has little effect on the estimates — if anything the sensitivity of the teenage years is even more pronounced. Lifting the age at which income is measured from 24 to 26, 28 or 30 also leaves the general conclusions unchanged.

C2. Family fixed effects

The key identifying assumption behind the methodology here, and in Chetty and Hendren (2018), is that selection effects do not vary with the age at move of the child. This seems unlikely to be true in a strict sense — certainly observables appear to differ slightly by age at move (Appendix Figure A.3) — but it remains unclear whether the extent of any variation is sufficient to meaningfully bias the results.

An obvious place to begin testing this assumption is through the addition of family fixed effects to control for any fixed differences between families moving with children at different ages. I also consider family-sex fixed effects given the evidence in Table 3 of heterogenous exposure effects by child sex. In these fixed effect tests, identification comes from comparing siblings in different cohorts who thus differ in both their length of time exposed to the destination (the e_m) and in the predicted outcomes of their destination relative to the origin Δ_{odps} (since these are allowed to vary by birth cohort s). This requires a greater degree of precision in the measurement of the predicted outcomes to avoid attenuation bias, so more stringent sample restrictions on the estimated precision in Δ_{odps} are also considered.

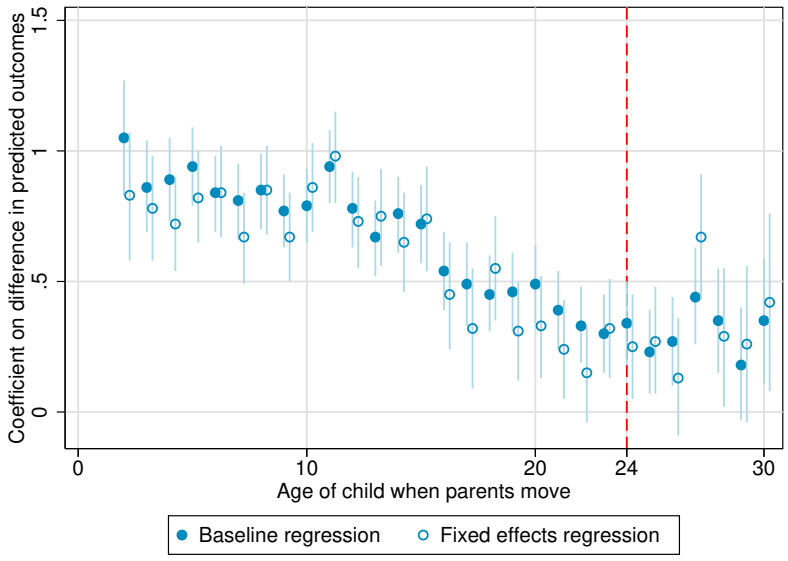


FIGURE C.1. PLACE EXPOSURE EFFECT ESTIMATES FOR CHILD INCOME RANK IN ADULTHOOD.

Notes: Estimated coefficients b_m from equations (2) and (3). The b_m capture the expected boost to an individual's household income rank at age 24 from moving at age m to a place with 1 percentile rank higher expected outcomes for permanent residents. They are estimated by regressing the adult ranks y_i of those whose parents move once in their childhood on the interaction of their age at parent move m with $\Delta_{odps} = \bar{y}_{dps} - \bar{y}_{ops}$ — the difference between the expected outcomes for permanent residents of the same parent percentile rank p and cohort s in the destination d versus the origin o. Controls vary across the specifications. Equation (2) includes indicators for cohort interacted with Δ_{odps} to capture potential mis-measurement of the origin capture, alongside fixed effects for each combination of parent income decile, origin, cohort and age at move. Equation (3) discards the fixed effects and includes instead: cohort and origin effects (via indicators for cohort and their interactions with predicted outcomes for permanent residents of the origin); and disruption effects (via indicators for age at move and their interaction with parental rank). This replicates Figure IV from Chetty and Hendren (2018).

The results are comforting. With family fixed effects, the estimated exposure effect falls modestly from 0.042 to around 0.03. With family-sex fixed effects, the fall is even less pronounced, with the estimates remaining at around 0.04. This suggests heterogeneity by child sex is important in the Australian setting. I also examine the selection effect — the expected boost to an individual's household income rank from having their parent move to a destination with 1 percentile rank higher expected outcomes *after* the child turns 24. With family-sex fixed effects this selection effect is halved and no longer statistically significant. It falls further towards zero as the sample is restricted to moves where the difference in origin and destination predicted outcomes is more precisely estimated.

Table C.1—Exposure effect estimates: more general specification and later ages of observation

	Baseline	General	Later age of observation		
	(1)	(2)	(3)	(4)	(5)
Early	0.011	-0.008	0.001	0.001	-0.013
	(0.007)	(0.009)	(0.012)	(0.023)	(0.052)
Late	0.042	0.052	0.044	0.045	0.044
	(0.003)	(0.005)	(0.005)	(0.006)	(0.007)
Post-outcome	0.008	0.013	-0.010	-0.004	0.033
	(0.013)	(0.019)	(0.015)	(0.016)	(0.021)
Age of observation	24	24	26	28	30
N	264,500	264,500	221,000	181,900	142,200

Notes: Estimates of the exposure effects $\gamma_{\tilde{m}}$ from equation 4 for early ($m \in \{2, ..., 11\}$), late ($m \in \{12, ..., 24\}$) or post-24 ($m \in \{25, ..., 34\}$) exposure, with either a more general set of controls (2) or for a later age of observation (3)-(5). These represent the expected boost to an individual's household income rank at the given age associated with an additional year at this stage of life in a destination with 1 percentile rank higher expected outcomes for permanent residents. They are estimated by regressing the adult ranks y_i of those whose parents move once in their childhood on the interaction of their time exposed to the destination at each life stage with $\Delta_{odps} = \bar{y}_{dps} - \bar{y}_{ops}$ — the difference between the expected outcomes for permanent residents of the same parent percentile rank p and cohort s in the destination d versus the origin o. In (1) and (3)-(5) controls capture: cohort and origin effects (via indicators for cohort and their interactions with predicted outcomes for permanent residents of the origin); disruption effects (via indicators for age at move and their interaction with parental rank); and indicators for cohort interacted with Δ_{odps} to capture potential mis-measurement of the origin. In (2) all but the last control is replaced by a much larger set of fixed effects for each combination of parent decile, origin, cohort and age at move. Murphy-Topel standard errors are in parentheses.

C3. Exogenous moves

A remaining concern is that there may be time-varying differences between families moving with children at different ages. Relationship breakdown, job loss or promotion could all give rise to moves, and themselves matter for outcomes in proportion to the time a child is exposed to them. The next test considers subsamples of moves that are more plausibly exogenous — moves out of locations in years with unusually large outflows for that location — and then re-estimates the exposure effects.

Let k_{pt} be the number of families leaving postcode p in financial year t as a proportion of the average number of families leaving the same postcode from 1991 to 2014. As in Chetty and Hendren (2018), many of those postcode-years with the highest relative outflows k_{pt} are associated with external shocks (such as mine closures in the Australian setting).[1] As noted by Chetty and Hendren (2018), while moves in subsamples with high values of k_{pt} may be more often for exogenous reasons, the destinations may still reflect endogenous choices. I follow them in instrumenting for Δ_{odps} and y_{ops} by $E[\Delta_{odps}|p, q]$ and $E[y_{ops}|p, q]$ — the

[1]Postcode-years with less than ten families leaving are dropped to avoid have high relative outflows that are driven by small underlying populations. I use the same threshold as in Chetty and Hendren (2018), purely to remain as close as reasonable to their specification.

Table C.2—Exposure effect estimates: family fixed effects

	Baseline	Family fixed effects			Family-sex fixed effects		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Early	0.011	-0.025	-0.012	-0.005	-0.021	-0.002	0.001
	(0.007)	(0.010)	(0.013)	(0.013)	(0.015)	(0.013)	(0.013)
Late	0.042	0.028	0.032	0.030	0.039	0.035	0.040
	(0.003)	(0.006)	(0.008)	(0.008)	(0.043)	(0.008)	(0.009)
Post-outcome	0.008	0.040	0.011	0.025	0.018	-0.036	-0.028
	(0.013)	(0.017)	(0.043)	(0.022)	(0.028)	(0.022)	(0.026)
Selection	0.292	0.365	0.293	0.287	0.140	0.097	0.047
	(0.068)	(0.104)	(0.145)	(0.133)	(0.361)	(0.123)	(0.140)
Sample s.e. on Δ_{odps}	< 2	< 2	< 1.75	< 1.5	< 2	< 1.75	< 1.5
N	264,500	263,100	228,300	175,400	263,100	228,300	175,400

Notes: Estimates of the exposure effects $\gamma_{\tilde{m}}$ from equation 4 for early ($m \in \{2, ..., 11\}$), late ($m \in \{12, ..., 24\}$) or post-outcome ($m \in \{25, ..., 34\}$) exposure, with either family or family-sex fixed effects. These represent the expected boost to an individual's household income rank at age 24 associated with an additional year at this stage of life in a destination with 1 percentile rank higher expected outcomes for permanent residents. They are estimated by regressing the adult ranks y_i of those whose parents move once in their childhood on the interaction of their time exposed to the destination at each life stage with $\Delta_{odps} = \bar{y}_{dps} - \bar{y}_{ops}$ — the difference between the expected outcomes for permanent residents of the same parent percentile rank and cohort in the destination versus the origin. Controls capture: cohort and origin effects (via indicators for cohort and their interactions with predicted outcomes for permanent residents of the origin); disruption effects (via indicators for age at move and their interaction with parental rank); and indicators for cohort interacted with Δ_{odps} to capture potential mis-measurement of the origin. Attention is restricted to families with five or fewer children. Murphy-Topel standard errors are in parentheses.

mean Δ_{odps} and y_{ops} for all movers in the sample from postcode p and in parental income decile q. I also present OLS estimates that do not account for endogenous choice of destination.

Figure C.2 shows the estimated late childhood exposure effect and its 95% confidence interval for subsamples drawn from moves that were part of progressively larger relative outflows from a postcode. I consider moves where k_{pt} was above its median value, 55^{th} percentile and so on to the 95^{th} percentile. The results are mixed. Below the 80^{th} percentile of relative postcode outflows the OLS exposure effect estimates are relatively close to the baseline estimate of 0.042. Beyond that point the estimates fall substantially, with negative point estimates and large standard errors for moves in the top decile of relative outflows. The IV estimates are more stable, but less precisely estimated. The average IV exposure effect estimate is 0.027, an attenuation of 30% relative to the baseline, with a less pronounced fall in point estimates in the top decile of relative outflows. In their (IV) estimates, Chetty and Hendren (2018) see a similar attenuation of around 20% on average, but if anything less attenuation of point estimates for the top decile.

Figure C.2 provides some comfort that the results are not driven by other factors correlated with *moderately* large relative postcode outflows, but the same cannot be said for the largest outflows. This validation exercise is thus less conclusive in the Australian setting than it appeared in the United States. One explanation

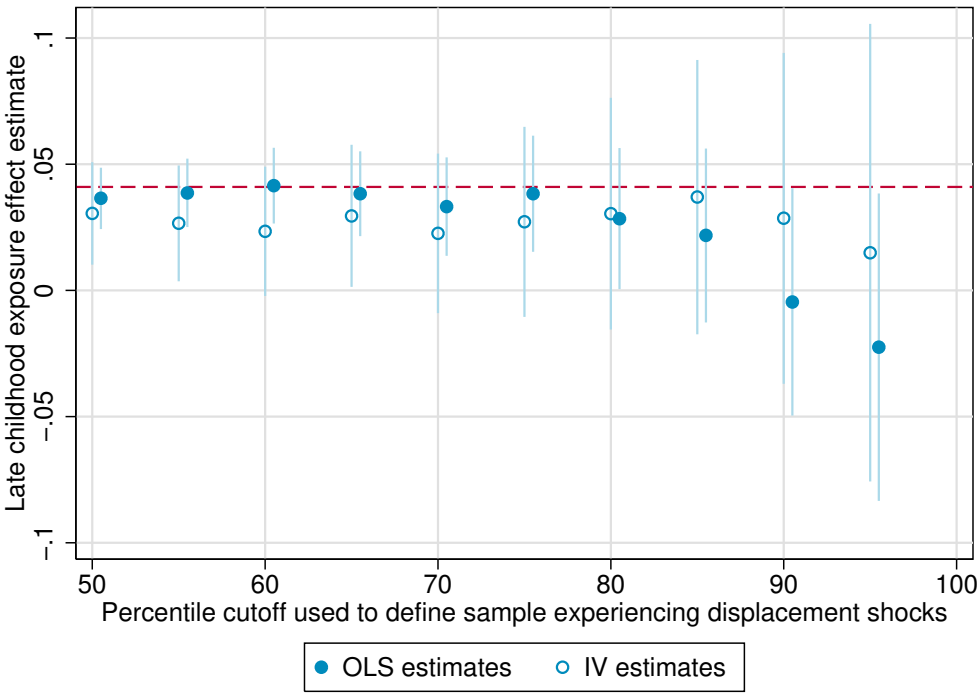


FIGURE C.2. PLACE EXPOSURE EFFECT ESTIMATES FOR PROGRESSIVELY LARGER DISPLACEMENT SHOCKS

Notes: Estimates of the exposure effects $\gamma_{\tilde{m}}$ and 95% confidence intervals from equation 4 for late childhood exposure, for subsamples of those moving out of postcodes in years with progressively higher relative outflows. These are identified by first calculating, for each postcode p and financial year t, the number of families leaving the postcode divided by the average annual number of families leaving the postcode from 1991 to 2014 (call it k_{pt}). Each individual in the 1-time mover sample is thus associated with a value of k_{pt} that indicates whether they were part of a relatively small $k_{pt} << 1$ or large outflow $k_{pt} >> 1$. The chart estimates the exposure effects for those with values of k_{pt} above its median value, its 55^{th} percentile and so on. OLS estimates are presented, alongside IV estimates where the origin and destination outcomes are instrumented for as described in the text. The IV estimates replicate Figure VI from Chetty and Hendren (2018).

is the failure of the identifying assumption — perhaps selection effects do vary with age. That said, the other validation exercises make this explanation more challenging to uphold. A more benign explanation may that the largest relative postcode outflows in Australia tend to be coupled with other factors that mitigate the effects of exposure to the destination. Indeed, fundamental differences in the treatment effects experienced by those choosing to move versus those forced to move are apparent in Chyn (2018).[2] This would be a threat to the external rather

[2]While Chyn (2018) finds larger treatment effects for those forced to move, this need not contradict the attenuation apparent in Figure C.2 if, as seems plausible, the appropriate specification of the treatment effect changes alongside its magnitude for exogenous shocks.

than the internal validity of the baseline estimates.

C4. Placebo tests

A final series of tests shows the outcomes of movers converges to those of permanent residents in a manner that picks up more than just the persistent differences in outcomes between the destination and origin. Rather, movers converge to the cohort- and gender-specific outcomes of permanent residents. Further, their outcomes mimic not just mean outcomes but the distribution of outcomes as well.

This greatly limits the potential for unobserved factors to explain away the exposure effects. For example, it seems unlikely that unobserved shocks when parents move — such as to income, wealth or family status — are correlated with the as-yet-unobserved *cohort-specific* predicted outcomes for a child. Such shocks seem far more likely to be correlated, if at all, with the persistent features of a place. Similarly, such shocks seem less likely to be correlated with gender- or distributional-specific features rather than the general features.

To begin, I show the best predictor of a mover's outcome is based on the experience of movers in their cohort, rather than those of surrounding cohorts. Following Chetty and Hendren (2018) I run two sets of regressions. In the first thirteen regressions I re-estimate the baseline specification in equation 4 as if an individual's financial year of birth was $s + l$ rather than s, where $l \in \{-6, ..., 6\}$. The resulting late childhood exposure effect estimates γ are in the solid dots in Figure C.3. Reflecting high serial correlation in a location's predicted outcomes, the exposure effects are all around the baseline estimate of 0.04. In the second single regression I re-estimate the baseline specification but include the lags and leads for the origin and difference terms. Where these lags or leads fall outside the sample window, the predicted outcomes are set to zero and an indicator I_l for the absence of that lag or lead is set to one. This gives rise to the specification below:

$$
\begin{aligned}
y_i = & \sum_{s=1978}^{1991} I(s_i = s)(\alpha_s^1 + \alpha_s^2 \bar{y}_{pos}) + \sum_{m=1}^{30} I(m_i = m)(\zeta_m^1 + \zeta_m^2 p_i) \\
& + \sum_{s=1978}^{1991} I(s_i = s)(\kappa_s \Delta_{odps}) \\
& + \sum_{l \in \{-6,-5,\ldots,5,6\}} \left(\sum_{\tilde{m} \in M} \delta_{\tilde{m}} + \gamma_{\tilde{m}} e_{\tilde{m}} \right) \Delta_{odp,s+l} \\
& + \sum_{l \in \{-6,-5,\ldots,5,6\}} \alpha_l \bar{y}_{po,s+l} + \omega_l I_l + \varepsilon_i
\end{aligned}
\tag{C1}
$$

The results are in the hollow dots in Figure C.3, and support a causal interpretation of the exposure effect estimates. The exposure effect estimate for the true cohort is only slightly attenuated. Further, while this estimate is statistically different from zero (with a p-value of 0.0057), the lags and leads are jointly insignificant (with a p-value of 0.20 on the joint test). It follows that any selection

process giving rise to the observed exposure effects must do so in a way that is correlated not just with the persistent features of a place, but its cohort-specific features — a more onerous requirement.

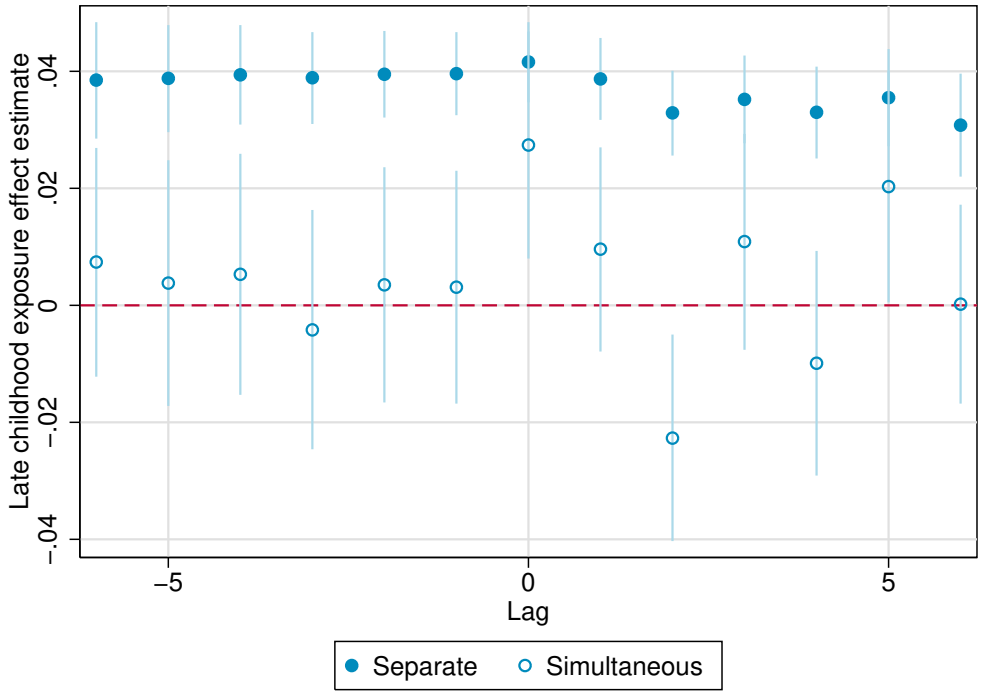


FIGURE C.3. PLACE EXPOSURE EFFECT ESTIMATES: EVENT STUDY

Notes: Estimates of the exposure effects $\gamma_{\tilde{m}}$ and 95% confidence intervals from equations 4 and C1 for late childhood exposure, when predicted outcomes are derived from a birth cohort that is not necessarily your own (solid dots) or when predicted outcomes for your birth cohort are included alongside those for neighboring cohorts (hollow dots). Thus the solid dots represent coefficients from thirteen separate regressions, using the predicted outcomes for those in financial year of birth cohort $s+l$ rather than an individual's actual birth cohort s, where $l \in \{-6, ..., 6\}$. The hollow dots run a single regression that includes the origin and difference in predicted outcome terms for all neighboring cohorts as in equation C1. Both these specifications allow for cohort effects (via indicators for cohort and their interactions with predicted outcomes for permanent residents of the origin); disruption effects (via indicators for age at move and their interaction with parental rank); and indicators for cohort interacted with Δ_{odps} to capture potential mis-measurement of the origin. This replicates Figure VII from Chetty and Hendren (2018).

Next, I show mover's outcomes converge to those of the same gender in their destination. Define by $\bar{y}^g_{pos}$ the predicted outcomes specific to a parent percentile rank, origin and cohort as before, but now also particular to a given gender g. Let the difference in predicted ranks between origin o and destination d be Δ^g_{odps}. We can now run three regressions, firstly equation 4 as before, but with $\bar{y}_{pos}$ and Δ_{odps}

replaced by $\bar{y}^g_{pos}$ and Δ^g_{odps} respectively. Next, we replace these key independent variables with the predictions for the opposite gender:

$$
\begin{aligned}
y_i^{80} = & \sum_{s=1978}^{1991} I(s_i = s)(\alpha_s^1 + \alpha_s^2 \bar{y}^g_{pos}) + \sum_{m=2}^{34} I(m_i = m)(\zeta_m^1 + \zeta_m^2 p_i) \\
& + \sum_{s=1978}^{1991} I(s_i = s)(\kappa_s \Delta^g_{odps}) \\
& + \sum_{\tilde{m} \in M} (\delta^*_{\tilde{m}} + \gamma^*_{\tilde{m}} e_{\tilde{m}}) \Delta^{-g}_{odps} + \alpha^3 \bar{y}^{-g}_{pos} + \varepsilon_i
\end{aligned} \tag{C2}
$$

and finally we include both the predictions for the true and opposite genders alongside one another:

$$
\begin{aligned}
y_i^{80} = & \sum_{s=1978}^{1991} I(s_i = s)(\alpha_s^1 + \alpha_s^2 \bar{y}^g_{pos}) + \sum_{m=2}^{34} I(m_i = m)(\zeta_m^1 + \zeta_m^2 p_i) \\
& + \sum_{s=1978}^{1991} I(s_i = s)(\kappa_s \Delta^g_{odps}) \\
& + \sum_{\tilde{m} \in M} (\delta^*_{\tilde{m}} + \gamma^*_{\tilde{m}} e_{\tilde{m}}) \Delta^{-g}_{odps} \\
& + \sum_{\tilde{m} \in M} (\delta_{\tilde{m}} + \gamma_{\tilde{m}} e_{\tilde{m}}) \Delta^{g}_{odps} + \alpha^3 \bar{y}^{-g}_{pos} + \varepsilon_i
\end{aligned} \tag{C3}
$$

The results from this exercise are in Table C.3. In column (1) we replicate our baseline specification using gender-specific predicted outcomes and see a similar late childhood exposure effect of 0.039. In column (2), and reflecting the fact that areas that are good for boys are typically good for girls as well, this exposure effect is only modestly attenuated when using predictions based on the opposite gender. However, when predictions for both own and opposite gender are included in the regression, the exposure effect is driven by the own gender predications with a coefficient of 0.037 (s.e. 0.008) versus 0.004 (s.e. 0.008). A remaining concern might be that families select into moves based on their child's gender in a way correlated with fixed family unobservables that matter for later child outcomes. To allay this concern, columns (4)-(6) repeat these regressions with family fixed effects and further restrict attention to families with both a boy and a girl in column (7). Again, the exposure effects are driven by the own gender predicted outcomes of place.

Finally, a similar exercise can be conducted by considering distributional rather than mean outcomes — for example, the event of falling into the top or bottom decile of the income distribution. In this case the three regressions compare the predictive power of the predicted outcomes for the true (distributional) outcome versus those for the mean outcome. The results from this exercise are in Table C.4. While predicted mean outcomes have some explanatory power over the probability an individual falls into the top or bottom decile, this disappears when conditioning on the predicted distributional outcomes. The outcomes of those

Table C.3—Late childhood exposure effect estimates: convergence in gendered outcomes

	Without family fixed effects			With family fixed effects			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Own gender prediction	0.039		0.038	0.037		0.040	0.041
	(0.005)		(0.008)	(0.011)		(0.013)	(0.016)
Opposite gender prediction		0.033	0.004		0.013	0.003	-0.009
		(0.005)	(0.008)		(0.011)	(0.013)	(0.016)
N	155,200	142,700	142,700	155,200	142,700	142,700	59,900

Notes: Estimates of the exposure effects $\gamma_{\tilde{m}}$ from variations of equation 4 for late ($m \in \{12, ..., 24\}$) exposure. These represent the expected boost to an individual's household income rank at age 24 associated with an additional year in late childhood in a destination with 1 percentile rank higher expected outcomes for permanent residents. They are estimated by regressing the adult ranks y_i of those whose parents move once in their childhood on the interaction of their time exposed to the destination in late childhood with $\Delta_{odps} = \bar{y}_{dps} - \bar{y}_{ops}$ — the difference between the expected outcomes for permanent residents of the same parent percentile rank and cohort in the destination versus the origin. Expected outcomes are based on permanent residents of the same gender (columns (1) and (4)); opposite gender (columns (2) and (5)); or both genders (columns (3), (6) and (7)). Controls capture: cohort and origin effects (via indicators for cohort and their interactions with predicted outcomes for permanent residents of the origin); disruption effects (via indicators for age at move and their interaction with parental rank); and indicators for cohort interacted with Δ_{odps} to capture potential mis-measurement of the origin. Family fixed effects are included in columns (3)-(7). Attention is restricted to families with five or fewer children, and the sample is further limited to families with children of both genders in column (7). Standard errors are in parentheses. This replicates Table IV from Chetty and Hendren (2018).

who move converge to those of the permanent residents of their destination not just in their mean outcomes, but in the distribution of their outcomes as well.[3]

C5. Summary

The results outlined in this section provide comfort as to the internal validity of the research design introduced in Chetty and Hendren (2018), both generally and in the Australian setting. Any unobserved factor explaining the observed exposure effects would need to operate within the family in proportion to time exposed and be able to replicate the cohort- and gender-specific outcomes of permanent residents, and the distribution of outcomes rather than just the mean. The examination of more plausibly exogenous moves leaves an important question mark over external validity, but is consistent with the finding of Chyn (2018) that

[3]In an earlier version of this paper I found the best predictors of a mover being in the top or bottom decile were based on the mean outcomes of the origin and destination Deutscher (2018). As noted there, this likely reflected the lower precision of the Australian predictions for permanent residents due to smaller geographic units and reduced geographic variation. The predicted probabilities of making the top or bottom decile are particularly imprecise and thus, if capturing more noise than signal, it is quite plausible that the predicted mean ranks may give a better indicator of the likely distributional outcomes of movers. In that version of this exercise I followed Chetty and Hendren (2018) and generated predicted distributional outcomes for each place as a quadratic (rather than linear) function of parent income rank. However, while this specification may more accurately capture nonlinearities in the tails of the relationship, it lowers the precision in the predicted distributional outcomes relative to predicted mean outcomes. In essence, we risk overfitting and generating noisier predicted distributional outcomes that thus have less predictive power over the outcomes of movers. In this exercise I generate predicted distributional and mean outcomes from simple linear regressions of the outcome on parent rank.

Table C.4—Late childhood exposure effect estimates: convergence in distributional outcomes

	Child in bottom decile			Child in top decile		
	(1)	(2)	(3)	(4)	(5)	(6)
Distributional prediction	0.029		0.020	0.045		0.035
	(0.005)		(0.008)	(0.003)		(0.006)
Mean rank prediction		-0.015	-0.006		0.033	0.010
		(0.003)	(0.005)		(0.004)	(0.006)
N	313,200	264,800	264,800	313,200	264,800	264,800

Notes: Estimates of the exposure effects $\gamma_{\tilde{m}}$ from variations of equation 4 for late ($m \in \{12, ..., 24\}$) exposure. These represent the expected boost to an individual's probability of ending up in the bottom or top decile of the income distribution associated with an additional year in late childhood in a destination with either a1 percentage point higher probability of the same outcome for permanent residents, or a 1 percentile rank higher mean outcomes for permanent residents. They are estimated by regressing the adult outcomes y_i of those whose parents move once in their childhood on the interaction of their time exposed to the destination in late childhood with $\Delta_{odps} = \bar{y}_{dps} - \bar{y}_{ops}$ — the difference in probabilities or mean outcomes for permanent residents of the same parent percentile rank and cohort in the destination versus the origin. Expected outcomes are based on the distributional outcomes (columns (1) and (4)); mean outcomes (columns (2) and (5)); or both (columns (3) and (6)). Controls capture: cohort and origin effects (via indicators for cohort and their interactions with predicted outcomes for permanent residents of the origin); disruption effects (via indicators for age at move and their interaction with parental rank); and indicators for cohort interacted with Δ_{odps} to capture potential mis-measurement of the origin. Standard errors are in parentheses. This replicates Table III from Chetty and Hendren (2018).

treatment effects may fundamentally differ between those choosing to move versus those forced to move.

Appendix D: Generated regressors, precision and valid inference

The equations estimated in this paper (and in Chetty and Hendren (2018)) fall into the more general class of two-step estimation, where regressors in the model of interest are generated from an auxiliary model. In particular, in the first step, the expected outcomes y for children born into a particular location l, cohort s, and parental household income rank p are predicted based on the sample of permanent residents of that location:

$$y_i = \alpha_{ls} + \beta_{ls} p_i + \varepsilon_i \tag{D1}$$

This model provides predicted values for the movers — denoted $\bar{y}_{ops}$ and $\bar{y}_{dps}$ — where we take their location l to be either their origin o or destination d respectively. Let $\hat{\beta}_1$ be the vector of estimated coefficients and X_{1o} and X_{1d} the matrices of observations indicating a mover's origin or destination respectively, along with their cohort and parent rank.

In the second step, these predicted values are used to generate regressors — $\bar{y}_{ops}$ and $\Delta_{odps} = \bar{y}_{dps} - \bar{y}_{ops}$ — for inclusion in a model for the outcomes of the movers:

$$\begin{aligned} y_i &= g(x_{2i}, \beta_2, \bar{y}_{dps}, \bar{y}_{ops}) + \varepsilon_i \\ &= g(x_{2i}, \beta_2, x_{1d}\hat{\beta}_1, x_{1o}\hat{\beta}_1) + \varepsilon_i \end{aligned} \tag{D2}$$

This is a classic example of the use of generated regressors. As noted in Pagan (1984), generated regressors pose a number of potential econometric issues. Perhaps most notably, while coefficients estimated from Equation D2 are generally consistent, the standard errors will not be, as they fail to account for uncertainty in the generated regressors. Perhaps reasonably, given they restrict attention to commuting zones with populations over 250,000, where the generated regressors are fairly precisely estimated, Chetty and Hendren (2018) do not consider this issue. However, given the Australian data is marked by smaller geographies and less geographic variation, this issue seems worth considering in more detail here.

D1. Valid inference

Murphy and Topel (1985) provide a procedure for calculating asymptotically correct standard errors in the fairly general circumstances. From the presentation in Greene (2003) the Murphy-Topel estimated covariance matrix for the model, given the two steps are estimated on different samples, is:

$$M = \hat{V}_2 + \hat{V}_2 \hat{C} \hat{V}_1 \hat{C}^T \hat{V}_2 \tag{D3}$$

where $\hat{V}_1$ and $\hat{V}_2$ are the estimated covariance matrices for models 1 and 2 respec-

tively and:

$$\hat{C} = \sum_{i=1}^{n} \left(\frac{\partial \ln f_{i2}}{\partial \hat{\beta}_2} \right) \left(\frac{\partial \ln f_{i2}}{\partial \hat{\beta}_1^T} \right) \tag{D4}$$

where f_{i1} and f_{i2} are the contributions of observation i to the likelihood functions of models 1 and 2 respectively. Now, we can follow the presentation in Hole (2006) and apply the chain rule to observe that:

$$\begin{aligned} \frac{\partial \ln f_{i2}}{\partial \hat{\beta}_2} &= \frac{\partial \ln f_{i2}}{\partial (x_{i2}\hat{\beta}_2)} \frac{\partial (x_{i2}\hat{\beta}_2)}{\partial \hat{\beta}_2} \\ &= \frac{\partial \ln f_{i2}}{\partial (x_{i2}\hat{\beta}_2)} x_{i2} \\ &= \frac{\partial \ln f_{i2}}{\partial \hat{y}_{mover}} x_{i2} \end{aligned}$$

and:

$$\begin{aligned} \frac{\partial \ln f_{i2}}{\partial \hat{\beta}_1} &= \frac{\partial \ln f_{i2}}{\partial (x_{i1o}\hat{\beta}_1)} \frac{\partial (x_{i1o}\hat{\beta}_1)}{\partial \hat{\beta}_1} + \frac{\partial \ln f_{i2}}{\partial (x_{i1d}\hat{\beta}_1)} \frac{\partial (x_{i1d}\hat{\beta}_1)}{\partial \hat{\beta}_1} \\ &= \frac{\partial \ln f_{i2}}{\partial (x_{i2}\hat{\beta}_2)} \left(\frac{\partial (x_{i2}\hat{\beta}_2)}{\partial (x_{i1o}\hat{\beta}_1)} x_{i1o} + \frac{\partial (x_{i2}\hat{\beta}_2)}{\partial (x_{i1d}\hat{\beta}_1)} x_{i1d} \right) \\ &= \frac{\partial \ln f_{i2}}{\partial \hat{y}_{mover}} \left(\frac{\partial \hat{y}_{mover}}{\partial \bar{y}_{ops}} x_{i1o} + \frac{\partial \hat{y}_{mover}}{\partial \bar{y}_{dps}} x_{i1d} \right) \end{aligned}$$

In both equations the first term is simply the score vector for model 2 — for simplicity denote its elements s_{i2}. The second equation includes derivatives in the brackets that simply pick up the estimated coefficients on the predicted values. The resulting estimate of $\hat{C}$ is as follows:

$$\hat{C} = X_2^T \,\mathrm{Diag}\left\{ s_{i2}^2 \frac{\partial \hat{y}_{mover}}{\partial \bar{y}_{ops}} \right\} X_{1o} + X_2^T \,\mathrm{Diag}\left\{ s_{i2}^2 \frac{\partial \hat{y}_{mover}}{\partial \bar{y}_{dps}} \right\} X_{1d} \tag{D5}$$

The above easily extends to the case where predicted values for neighboring cohorts are also included in the regression. The implementation of these standard errors in STATA has been outlined in Hardin (2002) and simplified in Hole (2006).

D2. Precision-based sample restrictions

Finally, throughout this paper, analysis is restricted in to those for whom the difference in predicted outcomes Δ_{odps} is more precisely estimated. The distribution of the standard error in Δ_{odps} for the 1-time movers sample is shown in Figure

D.1. For most of the analysis, I require $\Delta_{odps} < 2$, thus restricting attention to around the 80% of the sample for whom Δ_{odps} is most precisely estimated.

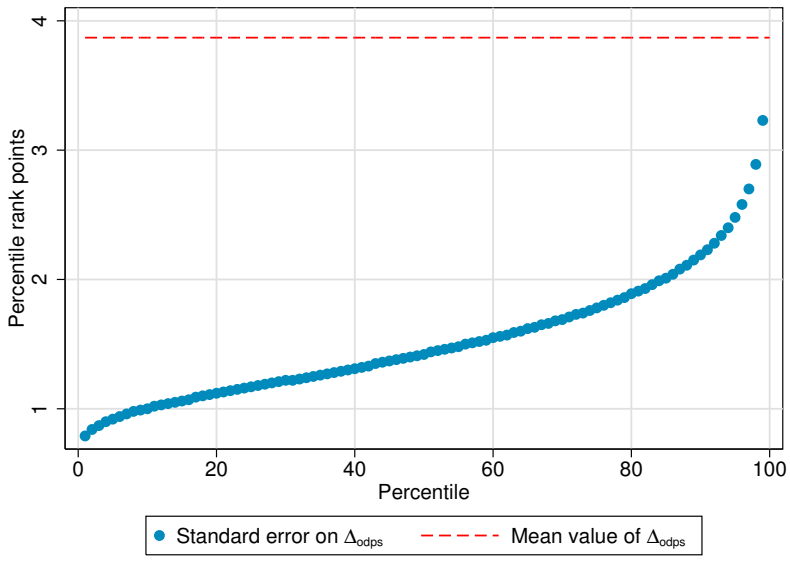


FIGURE D.1. DISTRIBUTION OF STANDARD ERROR IN DIFFERENCE IN PREDICTED OUTCOMES FOR PERMANENT RESIDENTS OF THE DESTINATION AND THE ORIGIN

Notes: For the 1-time mover sample, shows the distribution of the estimated standard errors on the key generated regressor: Δ_{odps}. Also shows the mean value of this regressor.

Key findings are robust to this precision-based sample restriction. In Table D.1, exposure effect estimates are shown for the baseline case, and for increasing levels of precision in Δ_{odps}. The results are not particularly sensitive to the choice of the precision-based sample restriction, with the late childhood exposure effect estimates all close to the baseline estimate of 0.042 and always larger than the early childhood exposure effect estimate.

TABLE D.1—EXPOSURE EFFECT ESTIMATES: VARYING LEVELS OF PRECISION IN Δ_{odps}

	Baseline	Increasing levels of precision				
	(1)	(2)	(3)	(4)	(5)	(6)
Early	0.011	0.006	0.008	0.011	0.030	0.010
	(0.007)	(0.006)	(0.006)	(0.007)	(0.009)	(0.023)
Late	0.042	0.039	0.040	0.042	0.040	0.046
	(0.003)	(0.003)	(0.003)	(0.003)	(0.005)	(0.013)
Post-outcome	0.008	-0.001	0.001	0.008	0.015	0.052
	(0.013)	(0.021)	(0.012)	(0.013)	(0.018)	(0.056)
Sample restrictions						
s.e. on Δ_{odps}	< 2	none	< 2.5	< 2	< 1.5	< 1
N	264,500	312,900	297,800	264,500	176,300	30,200

Notes: Estimates of the exposure effects $\gamma_{\tilde{m}}$ from equation 4 for early ($m \in \{2, ..., 11\}$), late ($m \in \{12, ..., 24\}$) or post-outcome ($m \in \{25, ..., 34\}$) exposure for the full baseline sample and larger or smaller samples based on varying restrictions on the standard error on Δ_{odps}. These represent the expected boost to an individual's household income rank at age 24 associated with an additional year at this stage of life in a destination with 1 percentile rank higher expected outcomes for permanent residents. They are estimated by regressing the adult ranks y_i of those whose parents move once in their childhood on the interaction of their time exposed to the destination at each life stage with $\Delta_{odps} = \bar{y}_{dps} - \bar{y}_{ops}$ — the difference between the expected outcomes for permanent residents of the same parent percentile rank and cohort in the destination versus the origin. the difference between the expected outcomes for permanent residents of the same parent percentile rank p and cohort s in the destination d versus the origin o. Controls capture: cohort and origin effects (via indicators for cohort and their interactions with predicted outcomes for permanent residents of the origin); disruption effects (via indicators for age at move and their interaction with parental rank); and indicators for cohort interacted with Δ_{odps} to capture potential mis-measurement of the origin. Murphy-Topel standard errors are in parentheses.

Appendix E: Intergenerational data construction

This Appendix describes the creation of the Australian Taxation Office's (ATO) de-identified intergenerational dataset. It is based on information provided by the ATO and those involved in the construction of the dataset.

E1. Overview

The dataset begins with the universe of federal tax returns from the 1991 to 2015 financial years, linked across individuals. This provides comprehensive information on individual incomes — the key challenge is linking parents and children.

Australia does not have two sources of parent-child links commonly used internationally. Birth register information is held by state and territories, and there is no national register as there is for Nordic countries. Further, parents are generally not required to provide identifying information for their children on tax returns, as family benefits are administered separately as cash transfers.[4] This rules out the methodology underlying Chetty et al. (2014), which uses the fact that parents' tax returns in the United States report their children's social security numbers.

Instead, parent-child links were formed by matching individuals to parents based on their reported residential addresses. Individuals report a residential address when they register for a tax file number — a unique personal identifier that is the closest Australian analogue to a social security number. The vast majority of individuals do this before they turn 17. These individuals are then linked to their likely parents based on residential addresses reported in tax returns. These links are disciplined by a set of more direct links available for a subset of individuals.

Address matching is behind the Statistics Canada dataset used in numerous widely-cited studies of intergenerational mobility (e.g. Corak and Heisz (1999); Oreopoulos (2003); Corak and Piraino (2011)). Yet the Australian institutional background, described below, means the ATO intergenerational dataset delivers a much higher match rate. Corak and Heisz (1999) report that they have parent links for around 49% of their selected Canadian cohorts. For the Australian cohorts studied in this paper, the link rate is around 92%, in line with that achieved by Chetty et al. (2014).

E2. Institutional background

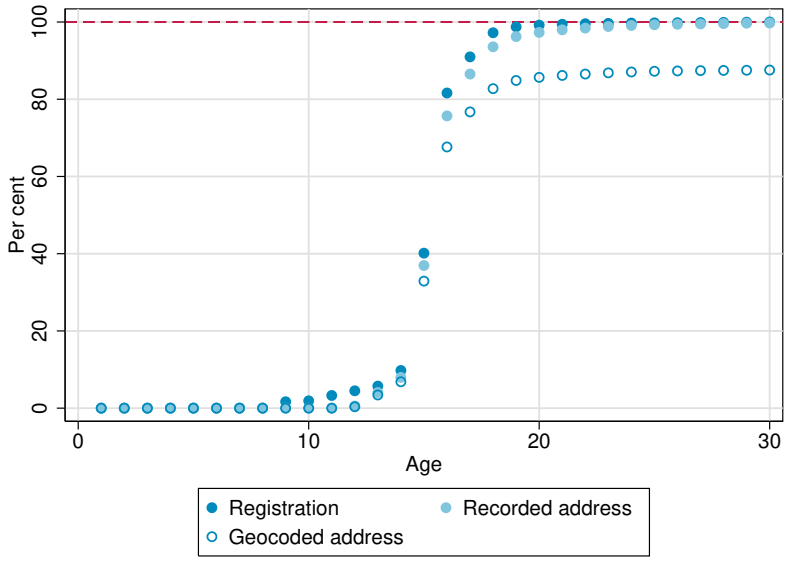
Address matching delivers high quality parent-child links in Australian tax data because most individuals register for a tax file number (TFN) with the ATO while still young and living in the family home. This reflects strong incentives to do so. Since its introduction in 1989, or shortly afterwards, a TFN has been needed to:

[4]Linking tax returns to this separate administrative database would have failed to provide complete parent-child links, as cash transfers have been and remain highly targeted, rather than universal.

- avoid paying higher withholding tax rates on labor and capital income;
- apply for unemployment, disability or family benefits; and
- apply for concessional loans for higher education.

As a result most transitions from childhood to independence — be it work, welfare or higher education — reward or require registering for a TFN. For example, of those born in Australia in the 1980 financial year and with a TFN by the time they were 30, over 90% had registered by age 17, and over 99% by age 20 (see Figure E.1). Importantly, a residential address is captured for most of these children at the point of registration, and is typically of sufficient quality to match to a geocoded address.

Figure E.1. Proportion registered for a TFN by age (Australia-born, 1980 birth cohort)



Notes: Darker blue dots show the proportion of registered clients born in Australia in the 1980 financial year who had registered by the given age. The lighter blue and hollow dots show the respective proportions with a recorded address and an address matched to a geocoded address by that age.

E3. Family linking procedure

The ATO dataset focuses on those born between the 1970 and 2000 income years (inclusive). Those born earlier are difficult to link to parents as many will have left the family home before the tax return panel begins in 1991. Similarly, many of those born later were yet to register for a TFN at the time the dataset was constructed.

Family links were generated for all individuals in the relevant birth cohorts, whether or not they were born in Australia. However, for the file used in this research, attention was restricted to those born in Australia. Country of birth is not directly observed in the tax data, but a good proxy for those born in Australia was derived based on other administrative information. From the 1978 birth cohort onwards this proxy performs particularly well, with the resulting Australian-born annual birth cohorts deviating by at most 1.5% from the population benchmark for the 1978-1991 cohorts.

E4. Family Tax Assistance links

Between the 1997 and 2000 income years, Family Tax Assistance (FTA) allowed low-to-middle income families to claim a higher effective tax free threshold. Low income families could claim the entire benefit through the payments system. However, middle income families had to provide the given names and dates of birth of their children on their tax returns. This provides a relatively direct source of family links for a subset of the child population.[5] These direct links then informed the algorithm for generating family links from the more widely available address links.

Initially, the details of all children a parent claimed between 1997 and 2000 were collected. This included a child's first name, date of birth and potential last names — while a child's actual last name is not listed, potential last names as inferred from those of their claiming parent and that parent's spouse. Duplicate claims were dropped and the remaining claims formed a base population of FTA children. FTA children were then linked to their adult selves among individuals registered for a TFN. A sequence of matches was performed, with only unmatched children passed to the next stage:

- *Perfect matches*: the first name, last name and date of birth match a unique individual;
- *First name error*: the last name and date of birth match a unique individual, where the two first names to have a levenshtein string edit distance of at most two;
- *DOB error*: the first name and last name match a unique individual, where the two years of birth are the same;
- *First name and DOB error*: the last name matches a unique individual, where the two first names have a levenshtein string edit distance of at most two and the two years of birth are the same;
- *Last name error*: the first name and date of birth match a unique individual.

[5]FTA claims do not necessarily imply a biological parent-child relationship, though in most cases the claimant will be a biological parent or primary carer.

Well over 70% of claimed children in each year were perfectly matched to an adult client. Fuzzy but unique matches allowed over 85% of claimed children in each year to be matched. These matches are spread across the birth cohorts of interest, with large numbers of those born from the 1980s onwards having FTA links.

E5. Address links

As FTA could only provide family links for a selective subset of the child population, the primary source of family links was based on shared residential addresses. As a first step, children were linked to all individuals who had ever lived at an address the child had lived at.[6] This forms the set of potential siblings and parents.

Siblings. — First, individuals were linked as siblings if:

- they had been at the same address within five years of one another;
- they both lived at that address before they turned 20;
- they had less than a 13 year age gap; and
- they had the same earliest last name.

These links were 'filled' out to ensure transitivity.[7] At the end of the parent linking process individuals were also linked as siblings if they shared the same parents.

Parents. — Individuals were then linked to parents. First, potential parents who were particularly young at the birth of the child (under 15 years of age) or old (45 years of age for women, 55 years of age for men) are dropped. Then the subsample of children who were *perfectly* matched as FTA children and have parent links as a result was isolated. A logistic regression was run on this subsample on the outcome that a potential parent is an FTA parent. The independent variables used in this regression were:

- potential parent sex interacted with an indicator for whether the potential parent and child share a last name;
- potential parent sex interacted with a quartic in parental age at birth either side of the median age at birth for that sex (29 for men, 27 for women)[8];

[6]The address was not required to be concurrent, as address histories in the tax data have gaps, and non-concurrent shared addresses in tax data may have been concurrent in reality.

[7]That is, if Alice is Bob's sibling and Bob is Charlie's sibling then Alice is Charlie's sibling. As a result children with more than a 13 year age gap may be identified as siblings if, for example, they share a sibling in common.

[8]Over the period 1975-1990, as calculated from Australian Bureau of Statistics (2017*b*).

- an indicator for whether the potential parent and child address histories imply they were at the same address at the same time;
- the length of overlap (in years) of a concurrent address episode;
- the distance (in years) separating a non-concurrent address episode; and
- the age of the child when first at the address (categorical between 13 and 25), interacted with whether the address episode was concurrent or not.

In the final step children were linked to their most probable parent, based on the logistic model's out-of-sample predictions for the probability a potential parent was an FTA parent. Each child was linked to the potential parent with the highest predicted probability of being an FTA parent, conditional on that probability being greater than 0.5. At the chosen threshold, less than 4% of the address-derived parents for the FTA subsample failed to match the FTA parent. Given the FTA links are not infallible, this seems reasonable. FTA link parents were then used for those children with no parent.

Postcode links. — Address matching is limited by the absence of complete address histories. Further back in the panel tax filers are less likely to have a recorded residential address. However, residential postcodes are reliably recorded — they are captured for the vast majority of tax filers in each of the years between 1991 and 2015. To exploit this, a set of supplementary links is based on residential postcode histories.

For all children, the postcode of their first address was extracted, typically their address when registering for a TFN. Children were also assigned their earliest recorded last name. Children were then linked to all individuals in the same postcode in the same year and with the same last name. This formed the set of potential parents. As in the address matching, potential parents who are particularly young at the birth of the child (under 15 years of age) or old (45 years of age for women, 55 years of age for men) were dropped. Given the large number of potential postcodes and last names, most children end up with a relatively small set of potential parents.

In the next step the subsample of children who perfectly matched FTA children and have parent links as a result is again isolated. Once again, a logistic regression was run on the binary outcome that a potential parent is an FTA parent. The independent variables used in this regression were:

- potential parent sex interacted with quartics in parental age at birth either side of the median age at birth for that sex (29 for men, 27 for women)[9];
- an indicator for if the potential parent's spouse is among the alternatives, also interacted with a categorical variable for the number of potential parents for the child (top coded at ten);

[9]Over the period 1975-1990, as calculated from Australian Bureau of Statistics (2017*b*).

- the age of the child when first at the postcode (categorical between 13 and 25).

In the final step children were linked to their most probable parent in the same manner as for the address matching. The exception is that here a slightly more conservative threshold was set — children were only linked to parents if the estimated probability of the potential parent being an FTA parent was greater than 0.75. The accuracy of this algorithm was only a little worse than the address matching. At the chosen threshold, only 8% of the supplementary parents for the FTA subsample fail to match the FTA parent.

Sibling links. — In this step, individuals with siblings were linked to the most probable parent for their family. The steps were as follows:

- look across groups of siblings (families);
- identify the most probable parent for each family — that is, the potential parent with the highest estimated probability of being a true parent[10]; and
- match individuals to the resulting most probable parent for their family.

Reassuringly, this process showed a great deal of consistency in the parent-child links. For children already linked to a parent, that parent is not replaced, or replaced by their spouse, in 90% of cases. Once children are matched to their most probable parent, those parents are matched to their earliest reported spouse over the period 1991 to 2015.

Resulting parent-child links. — The parent-child links resulting from this process are shown in Figure E.2. From 1978 to 1991 onwards the sample closely matches the size of the Australian-born population, deviating from the population benchmark by at most 1.5%. The proportion of the population linked to parents averages around 92% over this period as well. For the birth cohorts examined in this paper, 88% of links are derived from shared residential address, 4% from FTA, 2% from postcodes and 6% from siblings.

[10]FTA parents were assigned a probability of 1 and supplementary parents were all assigned to 0.75.

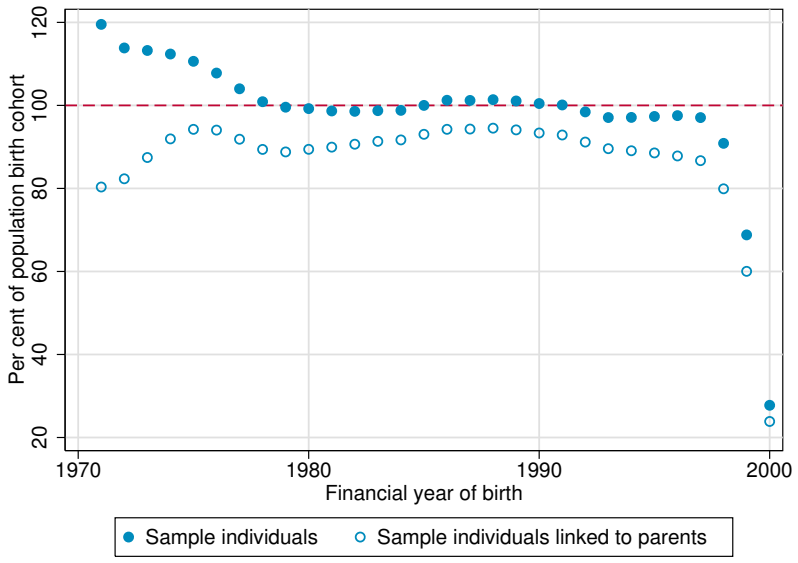


FIGURE E.2. SAMPLE COVERAGE RATES RELATIVE TO THE POPULATION OF INTEREST (%)

Notes: Shows the number of individuals in the sample, and the number linked to parents, as a percentage of the relevant population of interest. The population of interest is taken as the number of births in Australia (Australian Bureau of Statistics (2017*b*)), or for financial years prior to 1976 (where this data is not available) the estimated resident population aged zero on the last day (30 June) of the relevant financial year (Australian Bureau of Statistics (2017*a*)). Where both series are available they deviate by at most 2%.

*

REFERENCES

Australian Bureau of Statistics. 2017*a*. "Australian Demographic Statistics, June 2017." ABS Cat. No. 3101.0, Canberra.

Australian Bureau of Statistics. 2017*b*. "Births, Australia, 2016." ABS Cat. No. 3301.0, Canberra.

Black, Sandra E, Paul J Devereux, and Kjell G Salvanes. 2013. "Under pressure? The effect of peers on outcomes of young adults." *Journal of Labor Economics*, 31(1): 119–153.

Chetty, Raj, and Nathaniel Hendren. 2018. "The impacts of neighborhoods on intergenerational mobility I: Childhood exposure effects." *The Quarterly Journal of Economics*, 133(3): 1107–1162.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014. "Where is the land of Opportunity? The Geography of Intergenera-

tional Mobility in the United States." *The Quarterly Journal of Economics*, 129(4): 1553–1623.

Chyn, Eric. 2018. "Moved to opportunity: The long-run effects of public housing demolition on children." *American Economic Review*, 108(10): 3028–56.

Corak, Miles, and Andrew Heisz. 1999. "The intergenerational earnings and income mobility of Canadian men: Evidence from longitudinal income tax data." *Journal of Human Resources*, 34(3): 504–533.

Corak, Miles, and Patrizio Piraino. 2011. "The intergenerational transmission of employers." *Journal of Labor Economics*, 29(1): 37–68.

Deutscher, Nathan. 2018. "Place, jobs, peers and the teenage years: exposure effects and intergenerational mobility." Tax and Transfer Policy Institute Working Paper No. 10/2018.

Greene, William H. 2003. *Econometric analysis.* . 5th edition ed., Upper Saddle River, N.J.:Prentice Hall.

Hardin, James W. 2002. "The robust variance estimator for two-stage models." *Stata Journal*, 2(3): 253–266.

Hole, Arne Risa. 2006. "Calculating Murphy-Topel variance estimates in Stata: A simplified procedure." *Stata Journal*, 6(4): 521–529.

Murphy, Kevin M, and Robert H Topel. 1985. "Estimation and Inference in Two-Step Econometric Models." *Journal of Business & Economic Statistics*, 3(4): 370–79.

Oreopoulos, Philip. 2003. "The long-run consequences of living in a poor neighborhood." *The Quarterly Journal of Economics*, 118(4): 1533–1575.

Pagan, Adrian. 1984. "Econometric issues in the analysis of regressions with generated regressors." *International Economic Review*, 25(1): 221–247.