Carina Mood
Swedish Institute for Social Research (SOFI)
Work in progress: Included results are final, but more results may be added.

2010-12-03

Cautionary note: A tortuous overview only for those highly interested in the subject.

The importance of income definitions for the magnitudes and trends in intergenerational income correlations

Introduction

When studying the correlation between parents' and children's incomes, there are several choices to be made as regards the definition of income. These choices can be based on theoretical grounds, but can also depend on the availability and quality of data. In this memo, using data on Swedish men born 1960-1970, I show that such choices make a huge difference for the size of the estimated intergenerational income correlation (IIC) – in fact, with equally reasonable choices of the definition and coding of income the IIC can vary between 0.15 and 0.30. Furthermore, the decisions of how to define and measure income also impact on trends in IIC – I show that, for the cohorts mentioned, one can either find trendless fluctuations or a sizeable downward trend implying equalization.[1]

One lesson to be learned from the results that follow is that a single estimate of IIC is virtually useless unless the detailed construction of the income measure is explained; another that country estimates that are floating around in the international literature should be interpreted with great caution; and, finally, that international comparisons undertaken must take serious measures in order to achieve comparability. In the latter case, it is also of importance to note that estimates of IIC that represent a single or few years (or cohorts) may not represent an underlying level of inequality because these estimates depend also on period effects.

All analyses are done on the full population of Swedish men born 1960-1970 and whose parents are in Swedish registers. The data come from various registers, such as the register of the total population, the income register (based on tax records) and the multigenerational register (containing

[1] The intergenerational association can also be studied using regression coefficients (which are called elasticities if incomes for both parents and children are logged) instead of correlations, but in order to be able to isolate the influence of a range of factors on the association without making the results too complex, I here focus on correlations only. To limit the complexity I have also chosen to study the correlations only for sons.

links between parents and children). While the results based on these data are of obvious general interest, it is not as obvious that the details of the results would be found when using other types of data (e.g., from surveys), because one must then also consider different types of measurement error and non-response problems. Likewise, the impact of different choices of income definition may differ between countries and be more or less pertinent over time. For example, in my analyses choosing to include or exclude self-employed and farmers have little impact on IIC, but in countries where these social classes are more prominent this decision may be more consequential.

Different alternatives in defining intergenerational income correlations

When aiming to estimate the IIC, some of the important decisions in defining income are the following. The decisions concern parents' as well as children's incomes, and it is not always the case that one can and should use the same definition for both:

1. Should income be measured as household (normally equivalized) income or individual income?
2. At which ages should income be measured?
3. How many years should income be averaged over?
4. Which incomes should be included? Salary income is normally included, but one can include or exclude other incomes, such as incomes from self-employment, work-related benefits, other benefits and capital incomes.
5. How are zero incomes, negative incomes, and very high incomes to be treated?
6. Should incomes be transformed to logarithms? If yes, and if using income averaged over several years, should one take the average of log incomes or the log of the average income?

A fundamental distinction is the one between household and individual incomes. If the interest lies in the economic standard of living, household income is the best indicator, but if the interest lies in the individual income-generating capacity, individual income is the preferred measure. One can easily conceive of different combinations of parental and children income measures depending on the question motivating the study:

1. Does the economic situation during childhood and youth affect a person's economic situation as adult? This question is best studied by relating the disposable household income of parents during childhood/youth to the disposable household income of the child when adult. Children's disposable household income is the best indicator of the actual standard of living, but using it as the outcome variable means that the intergenerational correlation incorporates also mechanisms of eg. partner selection and homogamy.
2. Does the economic situation during childhood and youth affect a person's income-generating capacity? To study this, the ideal choice of variables is the disposable household income of parents during childhood/youth (as in 1) and the individual earnings or similar of the children when adult.
3. Does the income-generating capacity of one or both parents affect a person's income-generating capacity? (Earnings or similar of parents during childhood/youth – earnings or similar of child at adult age). This is the ideal model if we believe that parents' transmit (genetically or culturally) characteristics that enhance the income earning capacity.

Questions 1 and 2, using the economic situation during childhood, are the most relevant ones if one believes that parents with more money invest more in their children's human capital (primarily education), as in the Becker and Tomes (1979) theoretical framework. Even if one is primarily interested in the economic living standards of children (question 1), one may nevertheless find it better to use approach 2, as using the household income for children as the outcome variable means that the mechanisms behind the correlations are more complex and difficult to disentangle (though one can argue that spouse's incomes should be included because parents also invest in their children's marital opportunities by their choices of residential location and school). The fourth possible combination (earnings of parents and household incomes of children) appears of to be of less theoretical relevance.

Given the choice of household or individual income, it can generally be expected that the more reliable the income variable is (as an indicator of living standards or earnings capacity), the higher will the correlation be. This means that the inclusion of incomes that are likely to be less reliable, such as zero incomes, negative incomes and/or self-employment incomes is likely to suppress the intergenerational income correlation. Averaging incomes over several years also gives a more reliable estimate (Solon 1992; Corak and Heisz 1999), as does the choice of an age range for the observation of income that is likely to be more representative of the total income during the theoretical period of interest (sometimes, this period is the entire life of the parent, but it can also be e.g. the period when the child lives at home) (cf. Böhlmark and Lindqvist 2006).

It is a common finding, internationally (Bratsberg et al. 2007) and also in Sweden (Jonsson, Mood, and Bihagen 2010), that very high incomes are to a higher degree "inherited" than lower incomes. Top-coding or log-transforming the income variable means that the influence of the very high incomes is suppressed, and this should result in weaker correlations.

Previous results on correlations in Sweden are summarized in Table 1 (Appendix), and we can see that they vary widely both on the decisions made on the above points and in the magnitude of the intergenerational correlation. The IIC for sons ranges from 0.11 to 0.32, and the results from Jäntti et al (2006) and Österberg (2000) show clearly lower correlations than the more recent studies. The studies do however differ in so many respects that it is hard to judge which factors are driving the differences.

Results

To make the results easy to grasp, just one or two factors will be varied at a time. Throughout, I present results for both disposable income and labour income, because these two variables are so different that we cannot expect the conclusions for one of them to hold for the other. There is a set of default definitions that will all be subject to variations, although not all at once. If nothing else is stated, these are the reference values:

- The average real equivalized **disposable household income at age 8-15**, including only positive incomes, annual incomes >4 standard deviations are top-coded, and those with incomes from self-employment (including farming)>incomes from employment are excluded.
- The average real equivalized **disposable household income at age 33-37**, including only positive incomes, annual incomes >4 standard deviations are top-coded, and those with incomes from self-employment (including farming) > incomes from employment are excluded.
- The **father´s real labour income** (salary, self-employment income, work-related benefits) **at age 44-55**, including only positive incomes, annual incomes >4 standard deviations are top-coded, and those with incomes from self-employment (including farming) > incomes from employment are excluded.
- The average **real labour income** (salary, self-employment income, work-related benefits) **at age 33-37**, including only positive incomes, annual incomes >4 standard deviations are top-coded, and those with incomes from self-employment (including farming) > incomes from employment are excluded.

For all averages of incomes over years, it is required that at least three observations are non-missing and that it is not by definition impossible for the cohort/fathers in question to be observed some of the ages (e.g., for the cohort 1966, the average of income at ages 38-42 is not calculated, because there is no data on their income at 42). All analyses have been run selecting individuals with all observations in the interval non-missing, and if this makes a difference it is reported below.

Age when sons' incomes are measured

Figure 1 shows the correlation for different combination of income measures for the cohorts born 1960 to 1970, measuring child income at two age intervals: 33-37 and 38-42. Several interesting conclusions can be drawn from this figure. First, the correlation in labour incomes is much higher than the correlation in disposable incomes (around 0.3 as compared to around 0.2). Second, the correlation between the disposable income during childhood and the labour income when adult is close to the correlation in disposable incomes, suggesting that mechanisms of partner selection and family formation have only a minor impact on the intergenerational correlation.

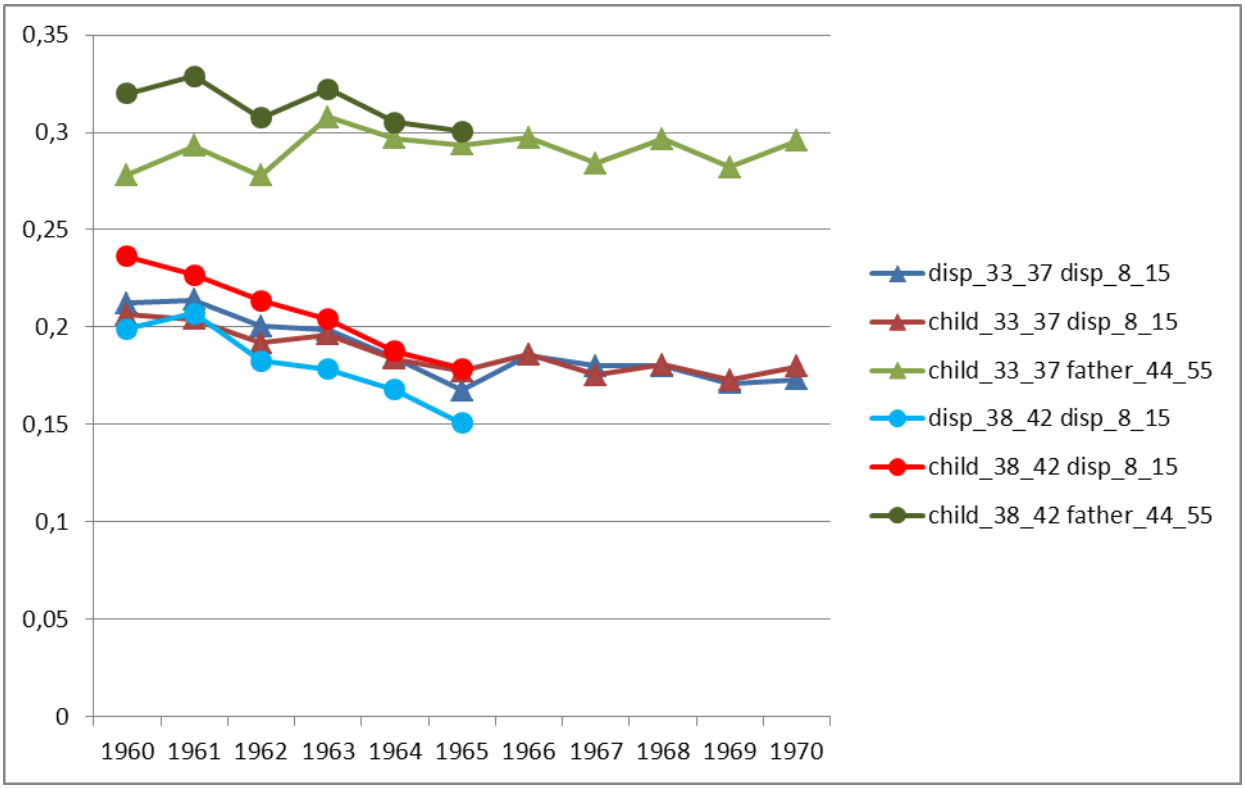

Figure 1. Intergenerational correlation for disposable family income and labour income, varying age interval when sons' income is measured.

Comparing the curves for ages 33-37 and 38-42 gives further information: For the correlation in labour incomes (the green lines), the sons' age range matters substantially for the cohorts born 1960-1963), but appears to matter little or nothing for the cohorts born 1964-1965. More detailed analyses, using also the intervals 34-38, 35-39, 36-40 and 37-41, show that the correlation for the oldest cohorts increases gradually with higher age intervals. Thus, the choice of age range for the sons matters not only for the magnitude of the correlation, but also for the trend in correlations: If we use the higher age intervals, we see a decreasing trend in the intergenerational correlation, but if we use the lower age interval, the trend is stable or even slightly increasing. This result may be caused by a period effect: For the older cohorts, the lower age interval coincide more with the 1990's recession than the higher age intervals, but for the younger cohorts none of the intervals occur during the recession. This suggests that we should be concerned not only with ages of measurement but also with the historical period when income is observed.

A similar pattern of differences between the age intervals for the older cohorts can be seen also when we study the relation between disposable income during childhood and labour income when adult (the red lines), but not when the adult outcome is disposable income. The correlation in disposable incomes during childhood and adulthood (the blue lines) follows the same pattern over cohorts whether we use the interval 33-37 or 38-42, but the magnitude of the correlation is somewhat higher for the younger age range. The likely explanation is that the disposable income reflects the labour income more strongly in younger ages, because less people have partners and children.

Number of income observations for sons

The intergenerational correlation in disposable incomes and labour incomes using different 5-year age intervals (33-37 and 38-42) for sons are reproduced in Figure 2, together with a 10-year interval spanning over the sons' ages 33-42. The results are rather striking: For labour income, the 10-year average produces a correlation almost identical to the average of income in ages 38-42, and for disposable household income the correlation is almost identical to the average of incomes in ages 33-37. This suggests that the younger age interval gives a better proxy for long-term disposable household income, but a worse proxy for long-term labour income.

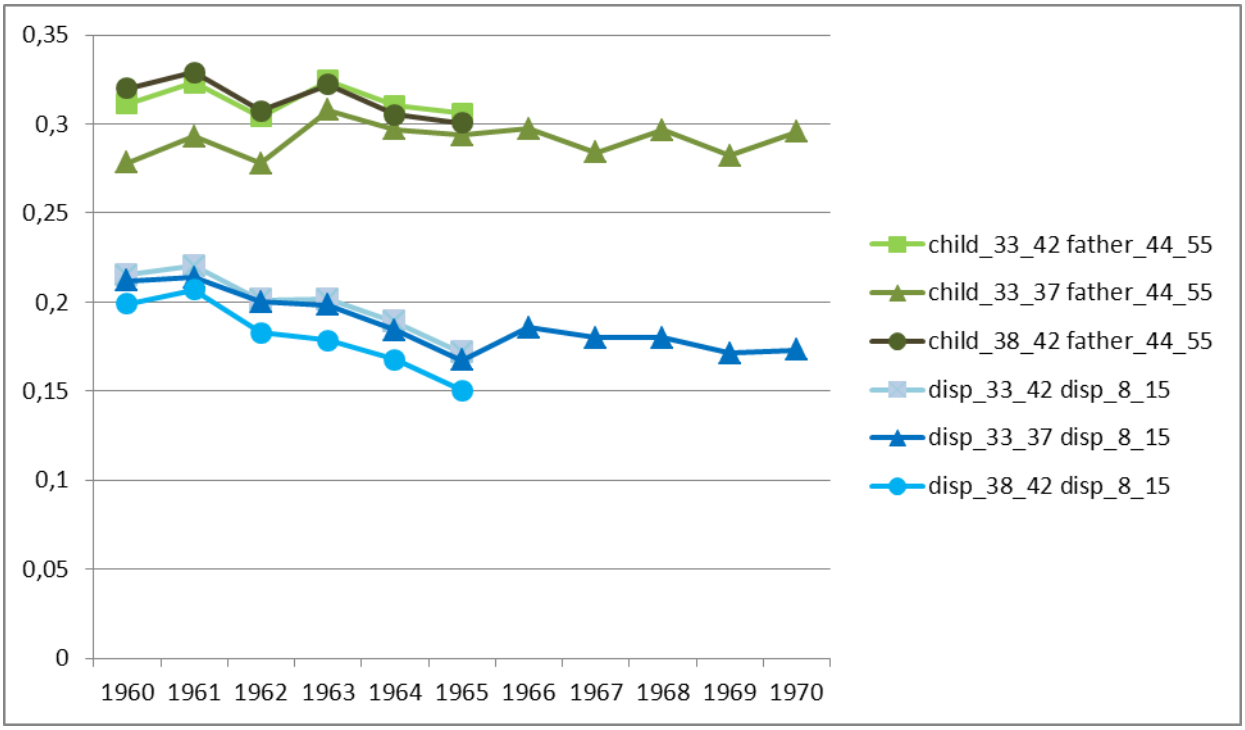

Figure 2 Income correlations using 10- (ages 33-42) and 5- (ages 33-37 and 38-42) year-intervals for sons' adult incomes.

Age when parental income is measured and number of observation for parental income

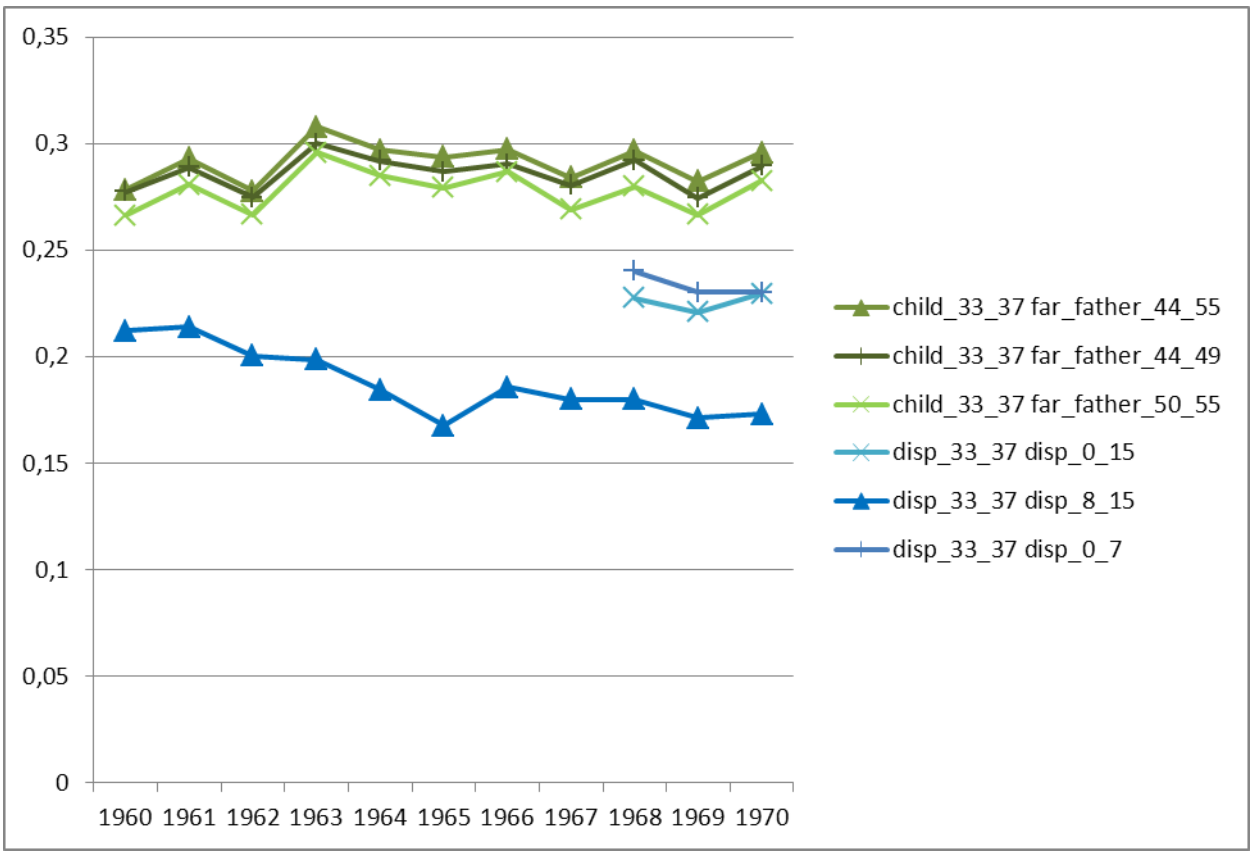

Figure 3 Income correlations using 12- (ages 44-55) and 6- (ages 44-49 and 50-55) year-intervals for fathers' incomes.

The first year that income data is available is 1968. A consequence of this is that studying fathers' incomes at young ages means that we have to select away fathers who had children relatively late in life (because there is no way of observing their incomes at young ages, i.e. before 1968). To avoid most of this selection, fathers' labour incomes are observed at rather high ages (44-55). In Figure 3, we contrast the correlations obtained for labour income with 5-year intervals (44-49 and 50-55) and a 10-year interval (44-55). We can see that the 10-year interval gives the highest correlation, though it is only slightly higher than the correlation using the 44-49 interval. The 50-55 interval gives a weaker correlation. These patterns hold regardless of whether we use the 33-37, 38-42 or the 33-42 interval for children's incomes (not shown in the figure). In additional analyses, I used younger intervals for fathers (thus excluding those who had children rather late in life), and the interval 40-44 gives about the same result as the 44-49 interval, but measuring fathers' income at ages lower than 40 gives correlations about as low as the one we get when using the 50-55 interval. Thus, it appears that the fathers' ages 40-49 result in the highest intergenerational correlation of labour incomes, and it does not matter much whether we use a five-year or a ten-year interval in this age span.

For the correlation in disposable incomes, the result is surprising: The correlation is much higher when using the average household disposable income when the child was of pre-school age (0-7) than for the age interval 8-15 (using the entire 0-15-year interval gives a slightly lower correlation

than when using the 0-7-interval). This result would suggest that the living standards in the very early years are more consequential for adult living standards, as found in a much cited US study (Duncan et al. 1998), but this does not appear intuitively correct for the Swedish case. There are two alternative explanations: 1. The living standard at older ages is equally important, but the disposable income when the child is older is a worse proxy for living standard (due to e.g. too heavy index weights of children in the equivalence scale) 2. Because the disposable income at the child's early years is more highly correlated with father's labour income, it is primarily the parents' income-generating capacity that is indicated, not the living standards. Further tests (not shown) reveal that the pattern holds also when studying only those who live with both biological parents during the entire 0-15 years period, so it is not caused by family type changes.

In Figure 4, the intergenerational correlations in equivalized disposable incomes (solid lines) are compared with the corresponding correlations using non-equivalized disposable income (dashed lines) in the father generation. Accounting for family size in disposable household income during childhood suppresses the intergenerational correlation *if childhood income is measured after pre-school age:* the intergenerational correlation is much stronger when using the non-equivalized disposable income. On the other hand, the correlations when childhood disposable income is measured at ages 0-7 or 0-15 remain unchanged.

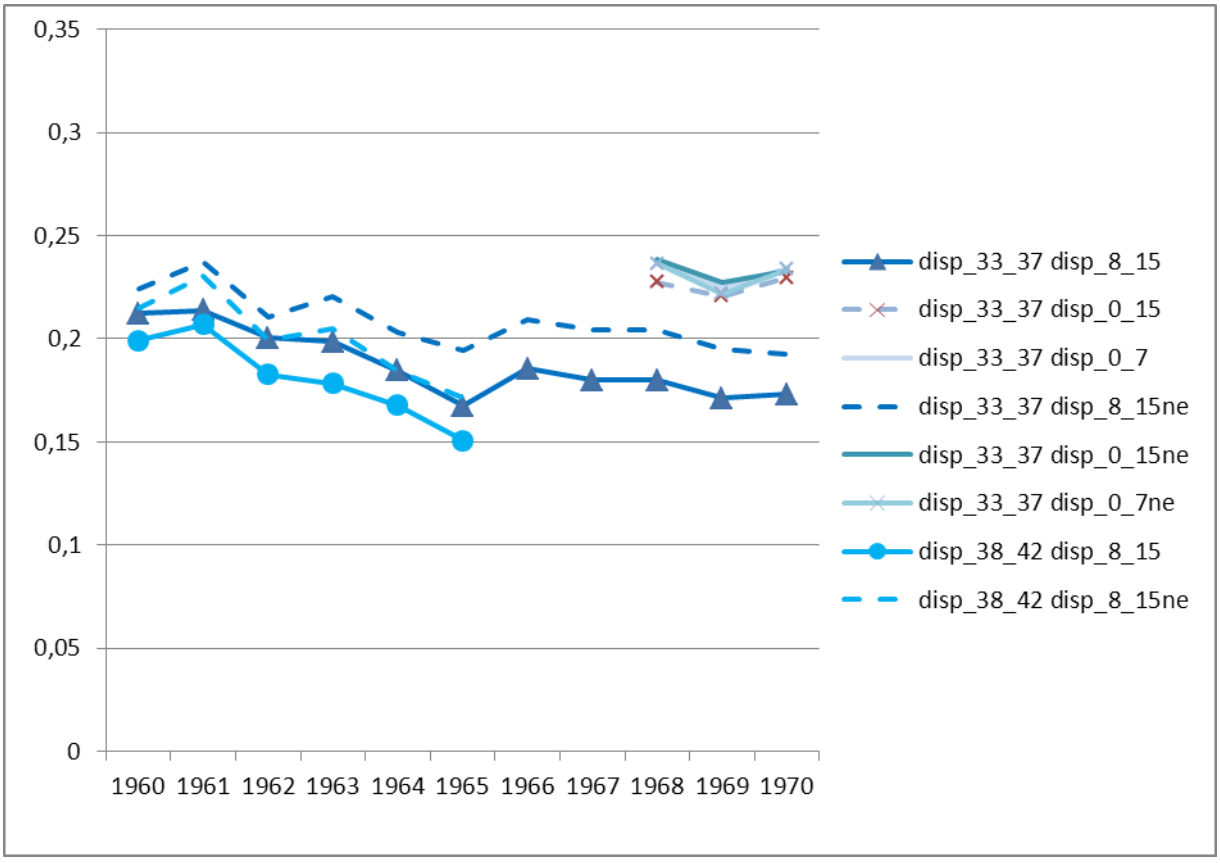

Figure 4 Intergenerational correlations, sons' disposable income at ages 33-37 and 38-42, and fathers' (1) disposable equivalized income (disp) (2) disposable non-equivalized income (disp_ne) and (3) fathers work income (fb), at children's ages 0-7 , 8-15 and 0-15.

When I replace disposable income with fathers' labour income for the same child age intervals (figure 5, orange lines), the correlations get even stronger, and we no longer see a stronger correlation if

father's income is measured when children are younger. The results of figures 4 and 5 suggest that father's income is more important than childhood living standards for children's incomes when adults, and that the correlation using disposable income during childhood will increase the more this variable correlates with father's work income.

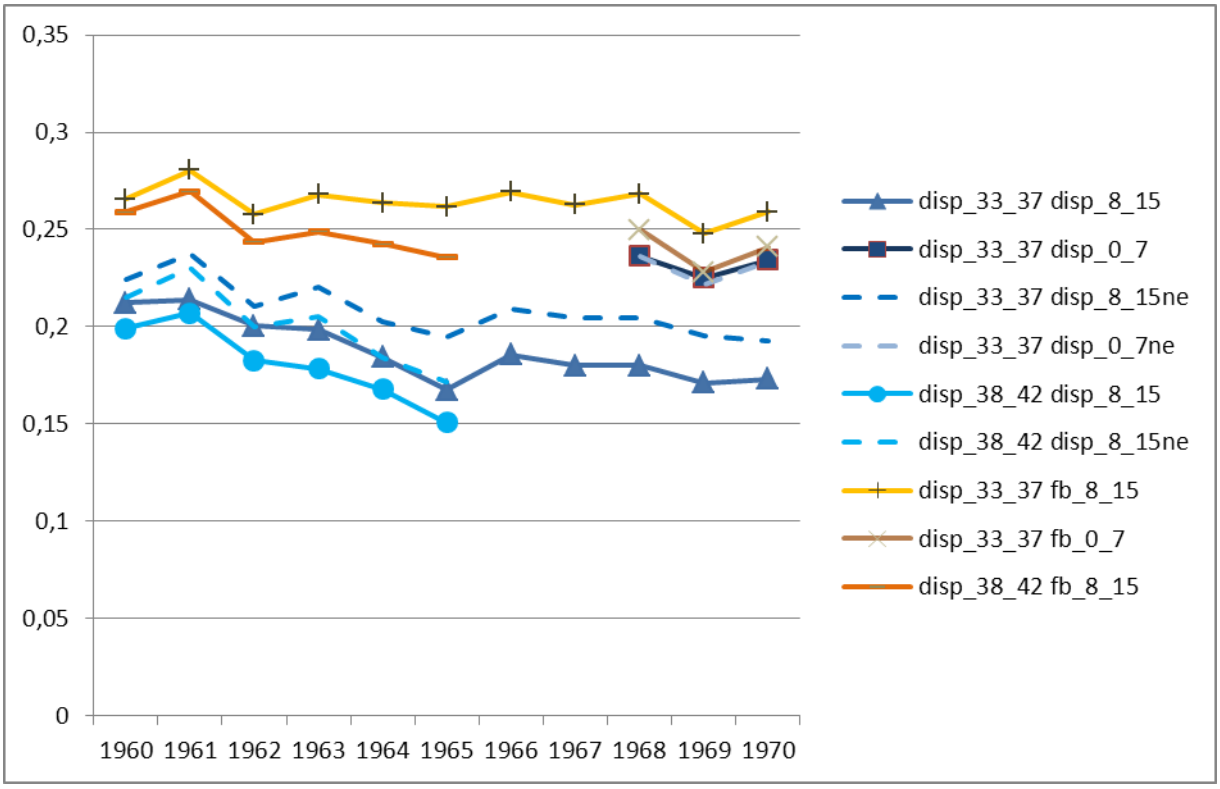

Figure 5 Intergenerational correlations, sons' disposable income at ages 33-37 and 38-42, and fathers' (1) disposable equivalized income (disp) (2) disposable non-equivalized income (disp_ne) and (3) fathers work income (fb), at children's ages 0-7 and 8-15.

[TO BE INSERTED: FURTHER ANALYSES OF NUMBER OF INCOME OBSERVATIONS FATHERS/SONS]

Inclusion of self-employed

When including fathers and sons who are self-employed (income from self-employment>income from employment), correlations remain very close to the original ones (**Figure 4**). This holds for both disposable income and labour income.

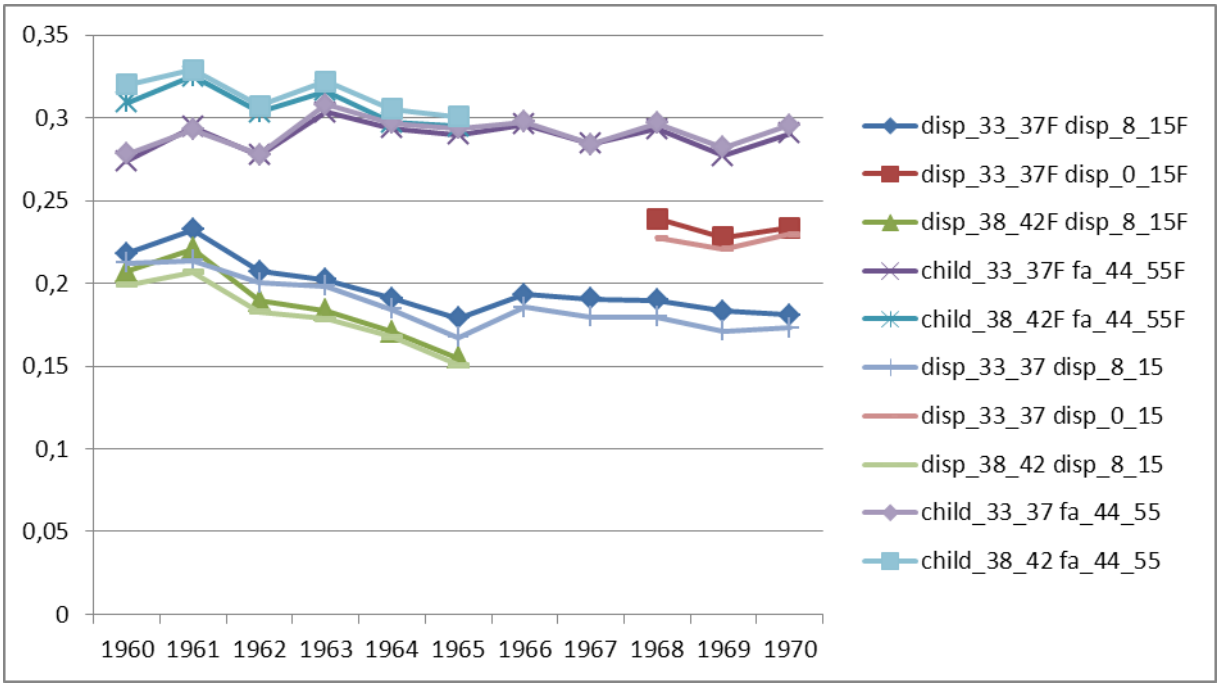

Figure 6 Income correlations including (subscript F) and excluding self-employed

Zero incomes, extreme incomes and logged incomes

In order not to complicate the presentation too much, the tests for the effect of excluding zero incomes, top-coding very high incomes, and using the log of incomes are shown only for labour income measured at fathers' ages 44-55 and children's ages 33 to 37 (the differences reported are the same if we instead use children's ages 38-42). Self-employed are excluded. The correlations are based on the same income definitions for father and son, so "no top-code" means that neither fathers' nor children's incomes are top-coded.

Figure 7 shows correlations for variables with different combinations of inclusion/exclusion of zeros and top-coding, and two different measures of log incomes. The log is taken of incomes without zeros and not top-coded. The results differ strongly between the different income definitions. The highest correlation is the default one seen in previous figures, which excludes zero incomes and top-codes the very highest incomes. Including zeros depresses the correlation somewhat, but contrary to expectations, not top-coding extreme incomes *lowers* the correlation substantially. Possibly, this is caused by some extremely influential outliers.

Using the log of incomes results in correlations that are much lower than the original ones (around 0.18 as compared to around 0.30). Papers using the log of income in intergenerational analyses seldom mention how the log of income is calculated. Two exceptions are Österberg (2000) and Hirvonen (2010), who both use the average of the log annual incomes (not the log of the average income). As it turns out, this alternative is the one that gives the lowest of all estimated correlations in Figure 7, and the difference in correlations when using this measure rather than the log of average incomes is quite large.

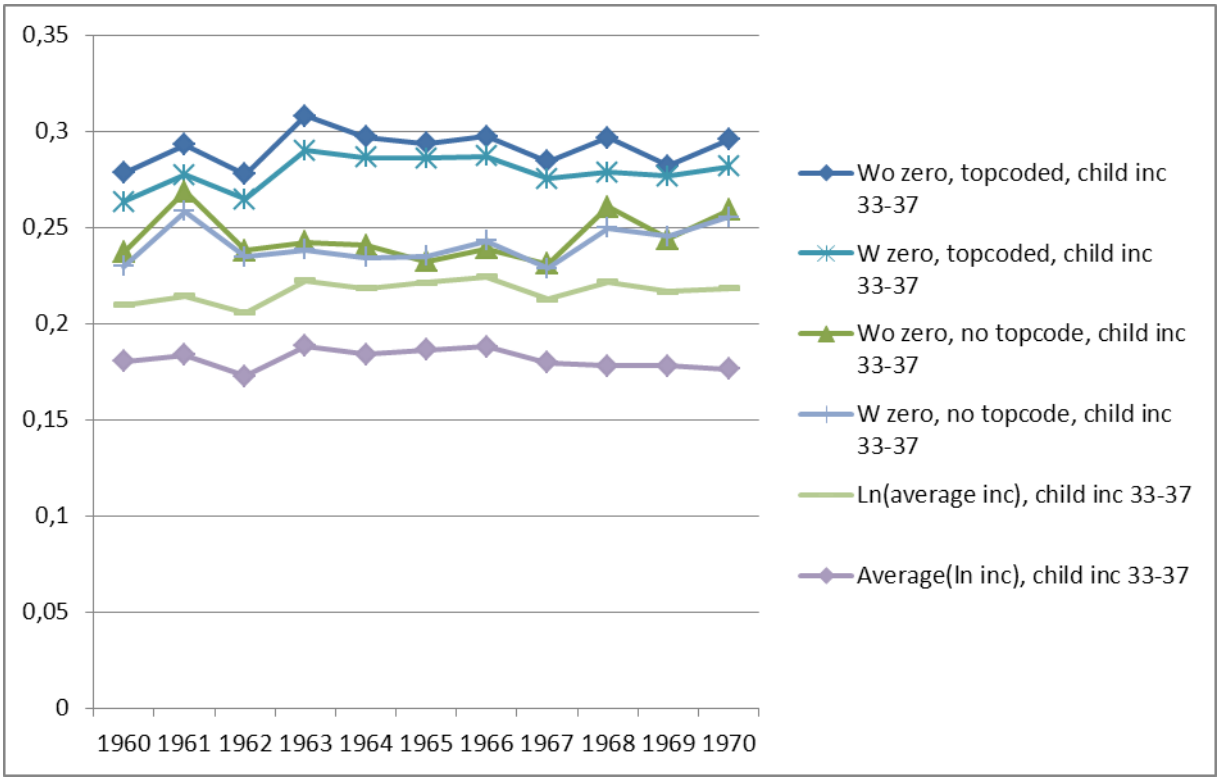

Figure 7. Income correlations for labour income with and without zero incomes and top-coding, and using two measures of log income.

Conclusions

In sum, the analysis of the intergenerational income correlations for eleven entire cohorts of Swedish men (1960-1970) shows that:

- The choice of studying labour income or disposable income is very consequential for the size of the correlation.
- The ages at which children's incomes are measured matter both for the levels and trends of the correlation.
- The importance of children's age on the correlation is different for disposable as compared with labour income.
- The number of income observations for sons and fathers is less important for the results than the ages of observation (though the minimum number of observations is five years).
- Including or excluding self-employed has little impact on the correlations.
- Including zero incomes suppresses the correlation
- Top-coding incomes leads to higher correlations

- Using the logarithm of incomes suppresses the correlations. This is particularly true if using the average of log incomes instead of the log of average incomes.

Naturally, there are many other possible variants of income definitions, and the tests here could have been (and may possibly be) elaborated in several ways. It is also quite possible that some patterns seen here may not hold in other countries, for other cohorts, or for survey data. Nevertheless, the results contribute to a systematic understanding of the variation in estimated intergenerational income correlations.

References

Becker Gary S. and N. Tomes. 1986. "Human Capital and the Rise and Fall of Families", *Journal of Labor Economics* 4/3: S1-S39.

Böhlmark Anders, and Matthew J. Lindqvist. 2006. "Life-Cycle Variations in the Association between Current and Lifetime Income: Replication and Extension for Sweden." *Journal of Labor Economics* 24: 879–96.

Bratsberg, Bernt, Knut Røed, Oddbjorn Raaum, Robin Naylor, Markus Jäntti, Tor Eriksson, and Eva Österbacka. 2007. "Nonlinearities in intergenerational earnings mobility: Consequences for cross-country comparisons". *Economic Journal* 117: C72–C92.

Breen, R, Jonsson JO, Mood C. 2010. *The role of social mobility for income mobility*. Manuscript.

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: 504–33.

Duncan, Greg J., W. Jean Yeung, Jeanne Brooks-Gunn and Judith R. Smith. 1998. "How Much Does Childhood Poverty Affect the Life Chances of Children?" *American Sociological Review* 63: 406-423.

Hirvonen, Lalaina. 2010. *Essays in Empirical Labour Economics: Family Background, Gender and Earnings*. PhD dissertation, Swedish Institute for Social Research, Stockholm University.

Jonsson, Jan O., Carina Mood, and Erik Bihagen. 2010. "Fattigdomens förändring, utbredning och dynamik." ("Poverty in Sweden: Recent Trends, Prevalence, and Dynamics") Chapter 3 (pp. 90-126) in *Social Rapport 2010*. Stockholm: Socialstyrelsen.

Jäntti, Markus, Bernt Bratsberg, Knut Røed, Oddbjørn Raaum, Robin Naylor, Eva Österbacka, Anders Björklund, Tor Eriksson. 2006. *American exceptionalism in a new light: a comparison of intergenerational earnings mobility in the Nordic countries, the United Kingdom and the United State.* IZA DP No. 1938

Mood, Carina, Jan O. Jonsson and Erik Bihagen . 2010 (forthcoming) *Socioeconomic persistence across generations: The role of cognitive and non-cognitive processes*. In: J. Ermisch, M. Jäntti, and T. Smeeding (eds.), Cross-National Research on the Intergenerational Transmission of Advantage. New York: Russell Sage.

Solon, Gary. 1992. Intergenerational Income Mobility in the United States. *American Economic Review (82),* 393-408.

Österberg, T. 2000 "Intergenerational income mobility in Sweden: What do tax-data show?" *Review of Income and Wealth* 46:421-436

Appendix: Table 1. Previous results on intergenerational correlations

	Data	Child cohorts	Child ages	Child income years	Parent	Parent income years	Sons/ daughters	Parents age	Income type parent	Income type child	N years parent	N years child	Excluded groups	Transform	Correlation
Jonsson et al 2010	Full register	60-70	33-37	93-07	Family	When child is 8-15 (1968-1985)	Both	mixed	Family disposable income	Family disposable income	8	5	self-emp, zero inc	Log	0,20-0,18
Mood et al 2010	Full register	62-65	38-42	00-07	Father	Ages 44-55, 68-07	Sons	44-55	Earnings +selfemp inc	Earnings +selfemp inc	12	5	self-emp, zero inc	None	0,31
Breen et al 2010	Full register	48-52	38-42	88-92	Father	68-72	Sons	mixed	Earnings +selfemp inc	Earnings +selfemp inc	5	5	self-emp, zero inc	None	0,32
Österberg 2000	Sample 1% register (SWIP)	-65	25- in 1990 (mean age in 1992=37)	90-92	Father	78-80	Sons	-64 in 1978 (mean age in 1978: 52)	Earnings	Earnings	3	3		None	0,18
Österberg 2000	Sample 1% register (SWIP)	-65	25- in 1990 (mean age in 1992=37)	90-92	Father	78-80	Sons	-64 in 1978 (mean age in 1978: 52)	Earnings	Earnings	3	3		Log	0,11
Österberg 2000	Sample 1% register (SWIP)	-65	25- in 1990 (mean age in 1992=37)	90-92	Father	78-80	Sons	-64 in 1978 (mean age in 1978: 52)	Earnings	Earnings	3	3	zero	Log	0,13
Österberg 2000	Sample 1% register (SWIP)	-65	25- in 1990 (mean age in 1992=37)	90-92	Father	78-80	Sons	-64 in 1978 (mean age in 1978: 52)	Market income (work+cap)	Market income (work+cap)	3	3		None	0,18
Österberg 2000	Sample 1% register (SWIP)	-65	25- in 1990 (mean age in 1992=37)	90-92	Father	78-80	Daughters	-64 in 1978 (mean age in 1978: 52)	Earnings	Earnings	3	3		None	0,13
Österberg 2000	Sample 1% register (SWIP)	-65	25- in 1990 (mean age in 1992=37)	90-92	Father	78-80	Daughters	-64 in 1978 (mean age in 1978: 52)	Earnings	Earnings	3	3		Log	0,07
Österberg 2000	Sample 1% register (SWIP)	-65	25- in 1990 (mean age in 1992=37)	90-92	Father	78-80	Daughters	-64 in 1978 (mean age in 1978: 52)	Earnings	Earnings	3	3	zero	Log	0,07
Österberg 2000	Sample 1% register (SWIP)	-65	25- in 1990 (mean age in 1992=37)	90-92	Father	78-80	Daughters	-64 in 1978 (mean age in 1978: 52)	Market income (work+cap)	Market income (work+cap)	3	3		None	0,14
Jäntti et al april 2006	Sample 20 % register	62	34-37	96-99	Father	75	Sons	35-64	Earnings, selfemp, sickness benefits	Earnings, selfemp, sickness benefits	1	4	zero	Log	0,13
Jäntti et al april 2006	Sample 20 % register	62	34-37	96-99	Father	75	Daughters	35-64	Earnings, selfemp, sickness benefits	Earnings, selfemp, sickness benefits	1	4	zero	Log	0,09