

Individual-level measures of skill specificity

This page contains links to the following information:

I. An explanation of the skill specificity measure, including variable definitions for different versions of this measure.

II. A Stata dataset containing six related measures of skill specificity at the individual level (see definitions below), which have been calculated for each wave of the International Social Survey Project (ISSP). The data include the original ISSP variables for survey ID, country ID, and individual ID (v1, v2, and v3 – see table below), so it is easy to merge the data with any particular ISSP survey. In contrast to Iversen and Soskice (APSR 2001), which used only ISSP-96 survey data, the measures in this data file have been calculated for all ISSP surveys and use OECD labor force statistics to calculate the labor force shares of each occupation. All calculations were made by Philipp Rehm (Duke and WZB) and are used in Cusack, Iversen and Rehm (2005).

If you use these data please cite [Cusack, Iversen and Rehm](#) (2005), as well as [Iversen and Soskice](#) (2001).

III. An Excel spreadsheet showing the calculation of the specificity measure, which can be used for any survey that includes information about occupation at the ISCO-88 2-digit level (or information that can be translated into this format – see IV below).

If you use the spreadsheet please cite [Cusack, Iversen and Rehm](#) (2005), as well as [Iversen and Soskice](#) (2001).

IV. ISCO conversion tables, translating ISCO-68 into ISCO-88, several national codes into ISCO-88, and higher-level ISCO-88 codes into lower-level codes. Once data are in ISCO-88 format, the Excel spreadsheet can be used to translate the codes into skill specificity values.

V. A note on the relationship between skills and class.

Explanation of the skill specificity measure

Relative skill specificity is a measure of how specialized an individual's skills are relative to the general skills or total skills that the same individual possesses. Using some simple notation, if s is a measure of specific skills and g is a measure of general skills, relative skill specificity can be defined as either s/g or $s/(s+g)$. The potential importance of relative skill specificity, defined either way, for explaining public policy preferences is modeled in Iversen and Soskice ([APSR 2001](#)) and in Iversen (*Capitalism, Democracy, and Welfare*, CUP 2005).

The measure is extracted from information contained in the ILO classification of occupations ([ISCO-88](#)) -- which uses the level and specialization of skills employed in an occupation to divide occupations into groups

(see [ILO/Hoffmann document](#)) – combined with OECD labor force data on the occupational distribution of employment. The exact procedure for calculating these measures is demonstrated in the Excel spreadsheet, and summarized in the table below. The theoretical logic behind the procedure is the following:

ISCO-88 is based on a hierarchical classification scheme. At the most aggregated level (major groups), occupations are identified by one of four levels of skills (corresponding to $s+g$), as well as a less systematic division into broad categories of jobs (there are 9 major groups in total). Each group is then broken into sub-groups based on the “the type of knowledge applied, tools and equipment used, materials worked on, or with, and the nature of the goods and services produced” ([ILO/Hoffmann](#), p. 6). At the most detailed level, there are 390 classes (called unit groups), each with a high level of homogeneity of skills. The number of unit groups in any higher level class will be a function of (a) the size of the labor market segment captured by that class, and (b) the degree of skill specialization of occupations found in that particular labor market segment. By dividing the share of unit groups by the share of the labor force in any higher-level group (using OECD Labor Force Surveys), we can generate a measure of the average skill specialization within that particular higher-level group. This measure corresponds to s .

For some purposes it may make sense to use s as a variable. The model for which we developed the measure, however, requires s to be divided by either total skills, $s/(s+g)$, or general skills, s/g . More precisely, the question we ask in the asset model of policy preferences is this: What happens to preferences as the balance of general and specific skills changes, holding income constant? (Iversen and Soskice 2001, 879). We use the ISCO-88 skill levels to approximate $(s+g)$, while the highest level of formal education is used to approximate g . If s instead of $s/(s+g)$ (or s/g) is used in a regression, the model is silent on the expected results. Or, more precisely, the prediction depends on the correlation between s and g , which the model says nothing about.

Finally, note that the skill measure is designed to capture *individual* heterogeneity, *not* national heterogeneity. If data are pooled across countries I recommend including country specific intercepts.

Variable descriptions

Variable	Values	Description
Skill Specificity variables: s1_lfs s2_lfs s_comp_lfs s1_lfs0 s2_lfs0 s_comp_lfs0	0 (low) to ca. 7. (If right censoring s_comp_lfs0 approximately at its 95 th percentile the max value is ca. 3.33)	Each variable measures an individual's relative skill specificity. As in Iversen & Soskice (2001), it is calculated as: [(Share of ISCO-88 level 4 groups)/(share of labor force)] divided by ISCO level of skills (s1) or highest level of education (s2), respectively (corresponding to $s+g$ and g). The _comp_ measures are calculated as an average of s1 and s2. We take the “share of labor force” from labor force surveys, as a grand mean over all country-years we have in the sample. Individuals not employed are left out of the lfs measures, while in the lfs0 measures they are coded 0.
v1 v2 v3 ccode yr_issp yr_field	1985-2000	za study number Respondent ID number Country ID Correlates of War country codes Year of ISSP survey (as it appears in title) Year of field work

A note on the relationship between skills and class.

A frequent question I get is “what is the relationship between the political economy model of skills, and more sociological models of class?” This is a simple question, but without a simple answer. I like to say that the asset model *is* a model of class. It has two key components: one is that skills determine the (expected) income of individuals and therefore their interest in redistribution; the other is that the specificity of those skills determine the (expected) costs of exogenous employment shocks, and therefore their preference for social insurance. But class also involves issues of collective organization (e.g., Korpi 1983; Stephens 1979), exposure to hierarchy and power relationships (e.g., Wright 1996), whether jobs involve “people” or “object” processing (Kitschelt 1994), etc. It can be objected to the asset theory that it does not capture these other dimensions of class, *as it does not*, and there is consequently an urge to test it against the alternatives. I welcome and encourage such tests. However, since the alternative conceptions of class often rely on simple classifications extracted from ILO’s standard classification of occupations (ISCO-88), which is also used in the construction of the skill variable, great care must be taken in designing the tests and interpreting the results. *This note explains the problems in using class dummies extracted from occupational classifications, and then suggests a remedy.* I hasten to add that this is not meant as a discussion of the hugely complicated concept of class (for that purpose, consult Wright 2005). Instead, it is a methodological note on a widely used operationalization of class.

To illustrate the key issues, I compare Svallfors’ (2004) application of Erikson and Goldthorpe’s (EG) well-known class scheme to the skill variable s_1 (which is defined as $s/(s+g)$). The EG classes rely on differences in the “logic of employment relations,” but the specifics need not concern us here (I have nothing of interest to say about the conceptual foundation of the EG scheme or any other class scheme – only the potential pitfalls in measuring and testing them against the asset model). There are six classes in the Svallfors adaptation of EG:

1. Service I
2. Service II
3. Routine non-manual
4. Skilled manual
5. Unskilled manual
6. Self-employed

With the exception of self-employed category (which in principle requires information outside the ISCO-88 classification), all classes are 0-1 dummies based on the most detailed level of ISCO-88 (two unit groups refer to self-employed, which, for our purposes, make up that class).

Now, imagine that someone wanted to carry out a strategic test of the asset model against the Erikson-Goldthorpe class model using $s/(s+g)$ versus the above class scheme. Such a test might be done using the following structural model:

$$P = a \cdot \frac{s}{(s+g)} + \sum b_i \cdot C_i$$

where P is the preference variable and C_i is the dummy for class i .

If we estimate this model and find that a is either negative or not significantly different from zero, while the class dummies perform as predicted, we might conclude that class and not skill specificity is what drives preferences. Assuming that no other variable (such as income!) matters, is such a conclusion warranted? Not necessarily – in fact, *not at all!*

The problem is that the explanatory variables are derived from the same classificatory scheme (ISCO-88), and therefore likely to be related. We can find out about this by defining the above variables on a simulated data set using the 390 ISCO-88 groups as “observations” (think of them as 390 individuals with different occupations). I’m grateful to Philipp Rehm for suggesting and implementing this approach. Thus we have one variable for s (explained above) one for $(s+g)$ (namely the ISCO skill levels), and one for each of the class dummies, and these are applied to the 390 “observations”. Are the variables related? Yes, of course, they are *all* related. But one relationship stands out: the multiple correlation coefficient for $(s+g)$ and the dummies is .90. In other words, the class dummies are closely related to skill level.

Let’s pursue this insight a bit further. First define $L = (s+g)$ and $L^* = f(C_i)$, where f is the estimated linear regression of L against C_i . Now the structural model is:

$$P = \alpha \cdot \frac{s}{L} + b \cdot L^*$$

Assume that the asset model is true. In fact, let’s assume that it determines P so that $P=s/L$ (the asset model also includes income as a determinant, but we can ignore that for our purposes). The $P=s/L$ assumption is convenient because we then know that whatever the regression results are, they will be consistent with the asset model. The question is whether we get results that correctly discriminate between the two models. As it turns out, we do not.

The problem can be easily illustrated if we imagine that $L^*=L$. In that case

$$\begin{aligned} \frac{s}{L} &= \alpha \cdot \frac{s}{L} + b \cdot L \\ \Downarrow \\ b &= (1-\alpha) \cdot \frac{s}{L^2} \end{aligned}$$

If we then substitute this expression for b in the structural equation we get:

$$P = \alpha \cdot \frac{s}{L} + (1-\alpha) \cdot \frac{s}{L}$$

In other words, α can be *any* number (negative or positive), and so, by implication, can b . If $L^*=L$, therefore, a regression of P on s/g and the class dummies is *not a strategic test at all*.

Now L^* is not quite the same as L (again $r=.90$), and perhaps the difference is systematically related to other aspects of class than skills. But who knows – considering how many different things the dummies could be capturing (they are just *dummies!*), the difference could be noise as far as the outcome of interest (P) goes. The denominator in s/L is less noisy – it is after all a deliberate attempt to measure skill level -- but that is not necessarily true for s/L since there is likely to be measurement error on s . Depending on the measurement error on the two variables, either a or b (or both) will pick up the effect of skill specificity. *In other words, with measurement error any result is consistent with either model.* Yikes, definitely not what we wanted!

But while a little measurement error can leave us completely in the dark (even with respect to the signs of the effects), there is a problem even *without* measurement error. To see this, imagine that s/L is the dependent variable. Since we don’t have a P variable in these data we can use s/L instead, just as an illustration. L and L^* are the independent variables. Obviously s/L and L are related by definition, and they are empirically as well as the following bivariate regression shows:

. reg s/L L

Source	SS	df	MS	Number of obs	=	388
Model	96.4375941	1	96.4375941	F(1, 386)	=	139.47
Residual	266.902664	386	.691457679	Prob > F	=	0.0000
Total	363.340258	387	.938863716	R-squared	=	0.2654
				Adj R-squared	=	0.2635
				Root MSE	=	.83154
s/L	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
L	-.5483868	.0464351	-11.81	0.000	-.6396842	-.4570893
_cons	2.993196	.1269053	23.59	0.000	2.743684	3.242709

Of course, since L and L^* are highly correlated, it is not surprising that s/L and L^* are also related:

Source	SS	df	MS	Number of obs	=	388
Model	114.607968	1	114.607968	F(1, 386)	=	177.86
Residual	248.732291	386	.644384172	Prob > F	=	0.0000
Total	363.340258	387	.938863716	R-squared	=	0.3154
				Adj R-squared	=	0.3137
				Root MSE	=	.80274
s/L	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
L^*	-.6648297	.0498512	-13.34	0.000	-.7628436	-.5668159
_cons	1.074923	.0556249	19.32	0.000	.9655576	1.184289

Note that the R-squared for L^* is in fact higher than for L . This is because L^* is also related to the numerator (s) in the dependent variable. The situation is analogous to one where L^* contains a component that is systematically related to P (other than the effect going through L).

But the response to this might be: “OK, whatever! We know that L is still related to s/L no matter what since we defined the variables so that they are related, right!?” Yes, but this is *not* what the multiple regression results actually tell us:

Source	SS	df	MS	Number of obs	=	388
Model	114.739745	2	57.3698724	F(2, 385)	=	88.85
Residual	248.600513	385	.645715619	Prob > F	=	0.0000
Total	363.340258	387	.938863716	R-squared	=	0.3158
				Adj R-squared	=	0.3122
				Root MSE	=	.80356
s/L	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
L	-.0471179	.1043006	-0.45	0.652	-.252188	.1579522
L^*	-.6175278	.1159915	-5.32	0.000	-.8455838	-.3894717
_cons	1.232285	.3527588	3.49	0.001	.5387097	1.92586

From these results we would erroneously conclude that L is unrelated to s/L . It is not simply that the statistical significance falls because multicollinearity pushed up the standard error on the coefficient on L (although it did do that), but that the size of the coefficient was reduced from $-.55$ to $-.05$. The effect of L^* , on the other hand, is largely unchanged. The results are similar if we used the class dummies in place of L^* . *But they are “wrong”!*

They are wrong in the sense that the dependent variable is related to L by definition, whereas the results indicate they are unrelated. The reason is that L^* is more closely related to s/L than L is, and since there is high collinearity between L and L^* , L will (loosely speaking) only capture what is not explained by L^* . One can think of L^* as capturing L plus something else that is relevant for s . The same can be true in a regression with P as the dependent variable and s/L and L^* as independent variables. If L^* captures s/L , plus something else that is relevant for explaining P , s/L may turn out to be insignificant (both statistically and substantively) *even though* it is in fact a critically important factor. *Skill specificity could be the headline story, but it is completely overlooked.* The morale of the story seems clear: *never use skill specificity and class variables derived from ISCO-88 in the same regression.*

Actually, the conclusion may be more far-reaching: using class dummies is not a very good way to test the effects of class. It is pretty obvious from what has been said that any analysis that uses EG class dummies as independent variables (or related class dummies based on ISCO-88) may simply pick up the effect of skill level. Now, if class is defined to capture differences in skill level (and perhaps the implied differences in expected income) then that's of course fine -- except that it is then far better to use a direct measure of skill level. If the class dummies are supposed to measure something other than skill level, well then it is certainly necessary to control for skill level. This never seems to be done. But even if it was, we know that the results do not necessarily make sense (as the above example illustrates).

What can be done? There is only one remedy, it seems to me. *Instead of measuring class by a set of dummies, devise a direct measure of the attribute of different occupations that the class concept is supposed to pick up.* Our concept of class identifies skill specificity as a key causal agent (in addition to income), and we try to measure it accordingly. If another causal agent is hierarchy, then measure the actual degree of hierarchy in different occupations. If it is a matter of “people-processing” versus “object-processing”, measure the nature of work by occupation along that dimension. If it is capacity for collective action, measure this capacity by occupation. And so on. Obviously, this makes the measurement task a lot harder, but the reward is getting results that are actually informative about the underlying causal mechanisms.

One example of this procedure is measuring occupational unemployment risk, which is used as an explanatory variable in Rehm (2005) and Cusack et al (2005). Since risks vary across occupations, one could simply create a set of occupational dummies and interpret their effects on preferences as a function of differences in unemployment risks. But that is likely to create the kind of problem of interpretation that is illustrated above. It is far better to measure actual unemployment risks (or at least unemployment) by occupation (as done in the Rehm paper).

References:

- Cusack, Thomas, Torben Iversen and Philipp Rehm. 2005. “Risk at Work: The Demand and Supply Sides of Government Redistribution.”
- Erikson, Robert and John H. Goldthorpe. 1992. *The Constant Flux: A Study of Class Mobility in Industrial Societies*. Oxford: Oxford University Press.
- Korpi, Walter. 1983. *The Democratic Class Struggle*. London: Routledge & Kegan Paul.
- Rehm, Philipp. 2005. "Citizen Support for the Welfare State: Determinants of Preferences for Income Redistribution." In Discussion Paper SP II 2005-02, Wissenschaftszentrum Berlin.
- Wright, Erik Olin. 1996. *Class, Crisis, and the State*. London: Verso.
- Wright, Erik Olin (ed.). 1996. *Approaches to Class Analysis*. Cambridge: Cambridge University Press.
- www.people. Stephens, John D. 1979. *The Transition From Capitalism to Socialism*. London: Macmillan.

Svallfors, Stefan. 2004. "Class, Attitudes and the Welfare State: Sweden in Comparative Perspective." *Social Policy & Administration* 38, no. 2 (2004): 119-38.

