A CAMSIS scale for the UK's Standard Occupational Classification 2020 and reflection on its 'multivariate validity'

Paul Lambert, University of Stirling

Presented to the Social Surveys and Social Statistics research group 'All Hands Seminar Day', Faculty of Social Sciences, University of Stirling, 4 December 2023

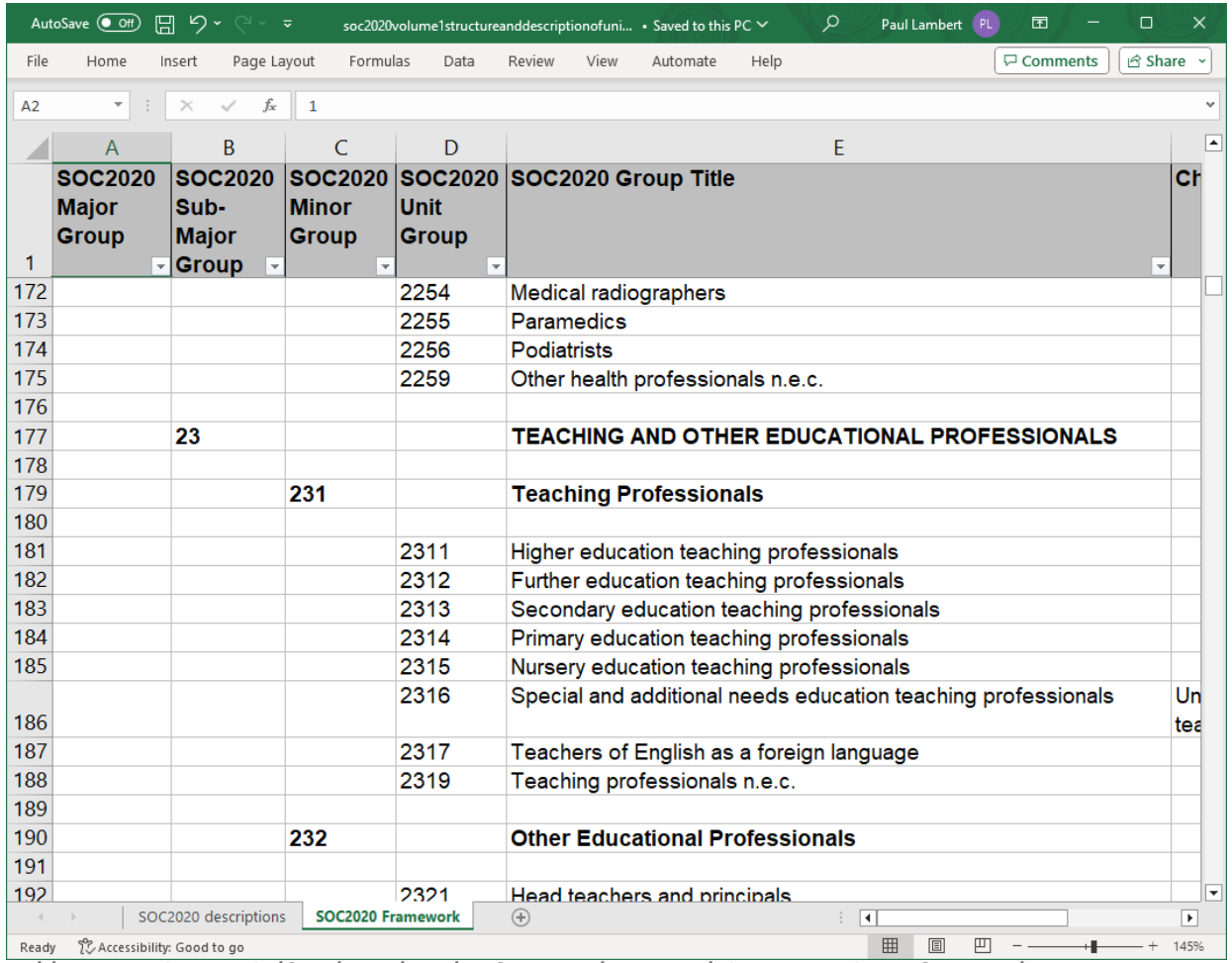
Abstract: CAMSIS scales - measures that are designed to be indicative of position in the structure of social stratification - can be generated by analysing data on the social interactions between the incumbents of occupational positions. Different scales can be generated for different taxonomies of occupational units, and for different populations of occupational incumbents. This analysis presents a CAMSIS scale generated for the UK's Standard Occupational Classification 2020 ('SOC2020'), based on analysis of UK Labour Force Survey data for 41261 both-working cohabiting heterosexual couples recorded 2021-23.

The process involved in deriving our recommended CAMSIS scale for SOC2020 ("gb_soc2020") raised an unexpected problem. Unlike in almost any previous comparable analysis, homogamy within 'professional occupations' seemed so persistent as to obscure differentiation between other occupational positions. We first describe the patterns observed using standard approaches, then present an adjustment to procedures that was felt to lead to an improved output.

Soon to distribute the gb_soc2020 measure (at www.camsis.stir.ac.uk), we also reflect on its properties. It has good validity in the conventional 'bivariate' sense. A more challenging question arises when multiple candidate measures of stratification position are added in combination to the same analytical procedure. In this context we could ask if measures demonstrate 'multivariate validity', that is, the capacity to disentangle different aspects or dimensions of stratification position compared to other measures. We seek to assess this systematically through appropriate Structural Equation Models, concluding that stratification measures have 'multiviarate validity' only when theorised as wide-ranging indicators.

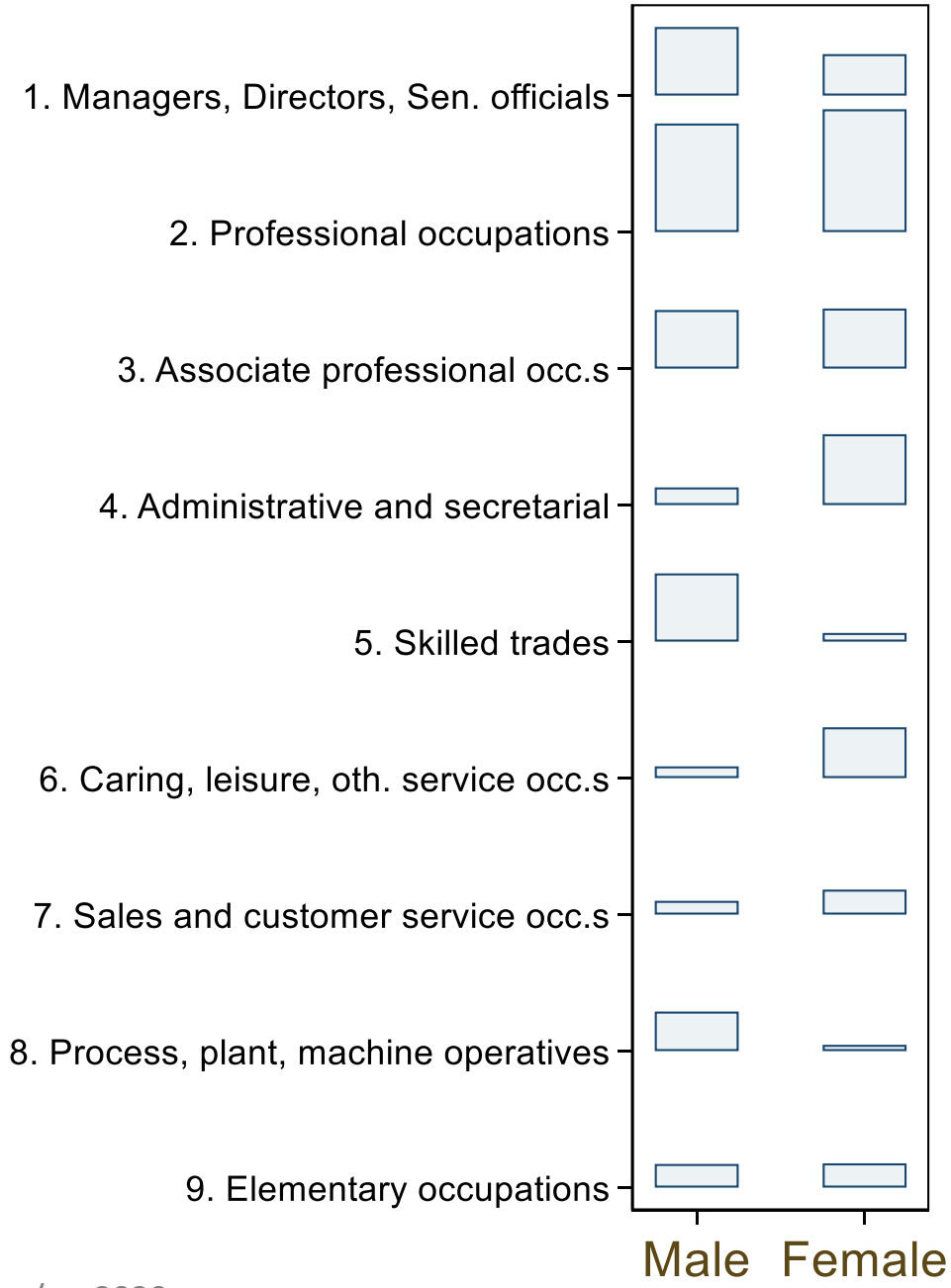
1) Standard Occupational Classification 2020

- Typically used on UK surveys from 2021 (3- or 4-digit version)
- Taxonomy of 412 4-digit units (valid range 1111-9269)
- Small modifications from SOC2000 and SOC2010 (steadily moving more cases into major group 2)

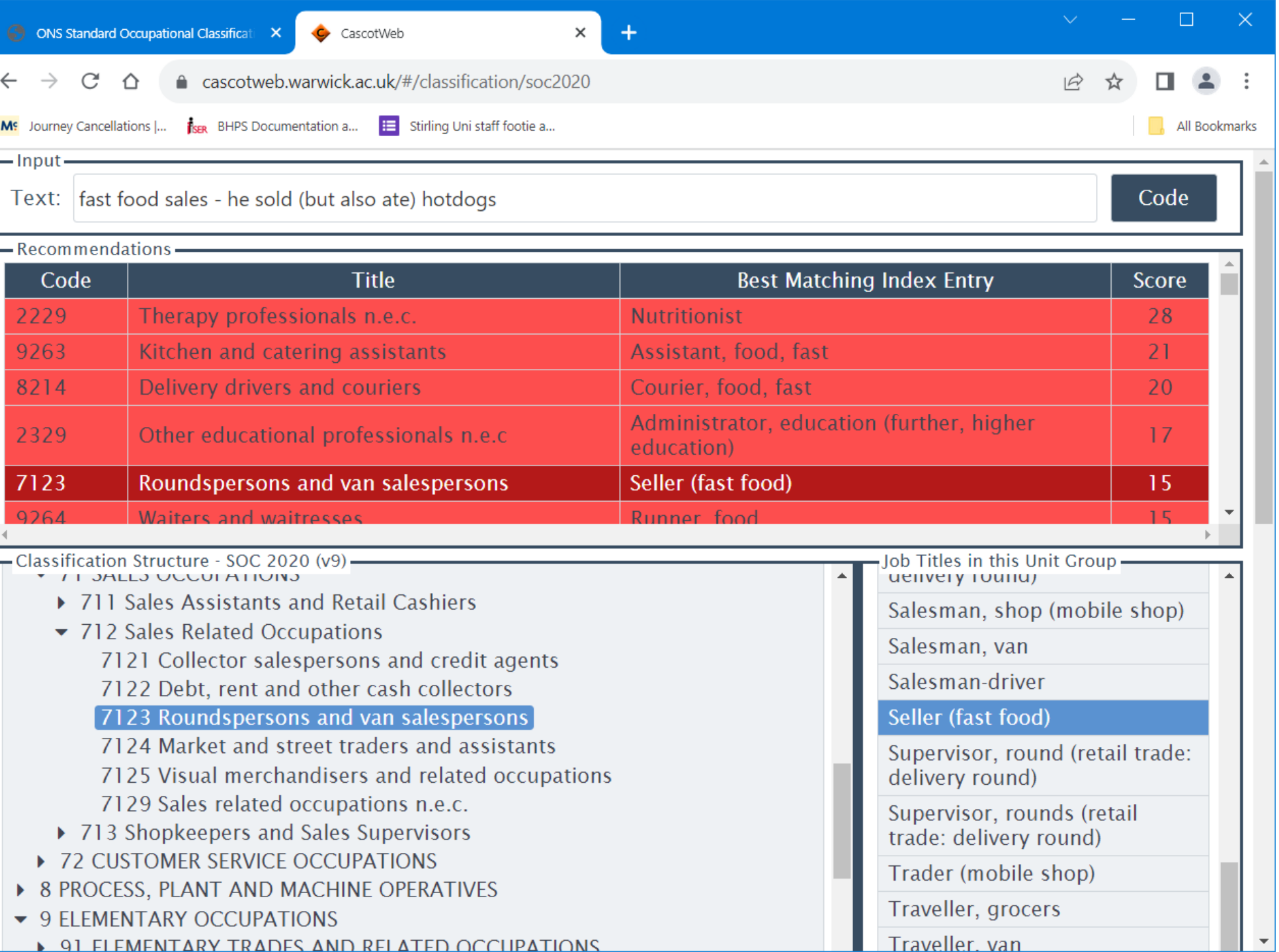


	A	B	C	D	E
1	SOC2020 Major Group	SOC2020 Sub-Major Group	SOC2020 Minor Group	SOC2020 Unit Group	SOC2020 Group Title
172				2254	Medical radiographers
173				2255	Paramedics
174				2256	Podiatrists
175				2259	Other health professionals n.e.c.
176					
177		23			**TEACHING AND OTHER EDUCATIONAL PROFESSIONALS**
178					
179			231		**Teaching Professionals**
180					
181				2311	Higher education teaching professionals
182				2312	Further education teaching professionals
183				2313	Secondary education teaching professionals
184				2314	Primary education teaching professionals
185				2315	Nursery education teaching professionals
186				2316	Special and additional needs education teaching professionals
187				2317	Teachers of English as a foreign language
188				2319	Teaching professionals n.e.c.
189					
190			232		**Other Educational Professionals**
191					
192				2321	Head teachers and principals

SOC 2020, 2021-23



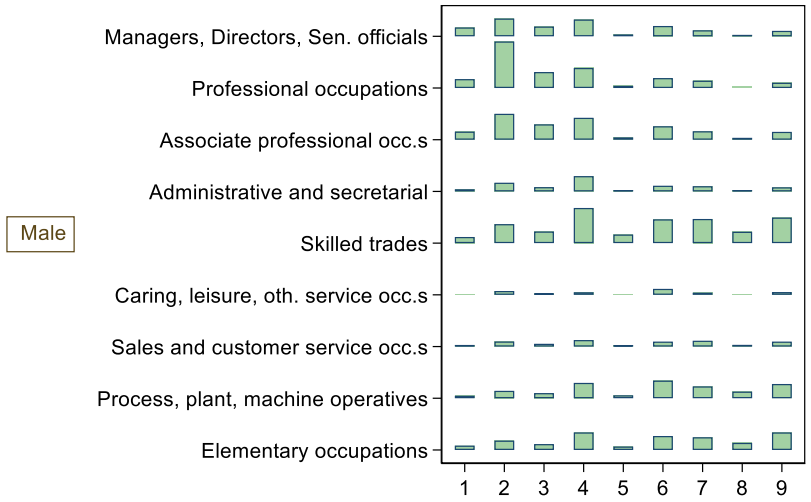
https://warwick.ac.uk/fac/soc/ier/software/cascot/choose_classificatio/
https://www.ons.gov.uk/methodology/classificationsandstandards/standardoccupationalclassificationsoc/soc2020



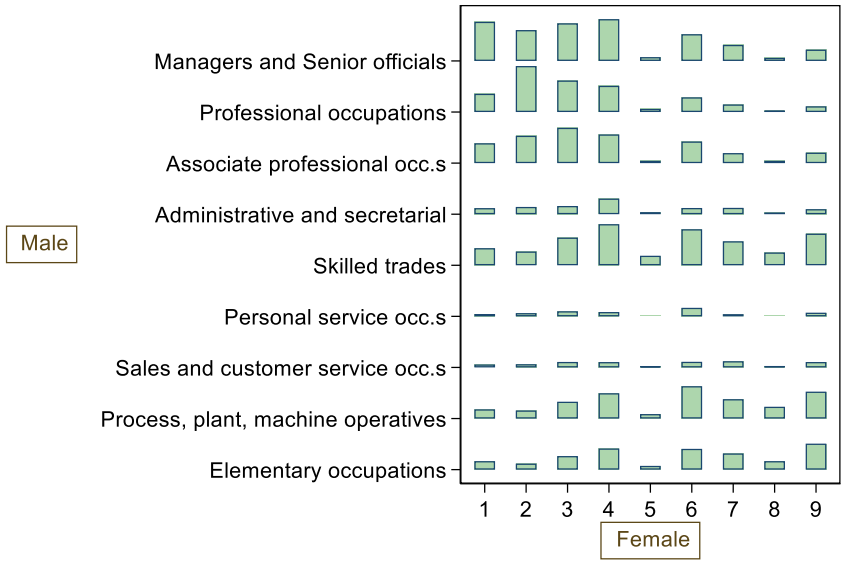
Occupational information reminders...

- Occupational unit group codes calculated by text coding
- Designed for 4-digit detail though many analysts use 3-, 2- or 1-digit versions
- Funded surveys also collect employment status data (self-employed with lots, few or no employees, manager of lots or of few, supervisor, employee) which may or may not feature in social classifications
- Metadata available to link SOC codes to stratification measures

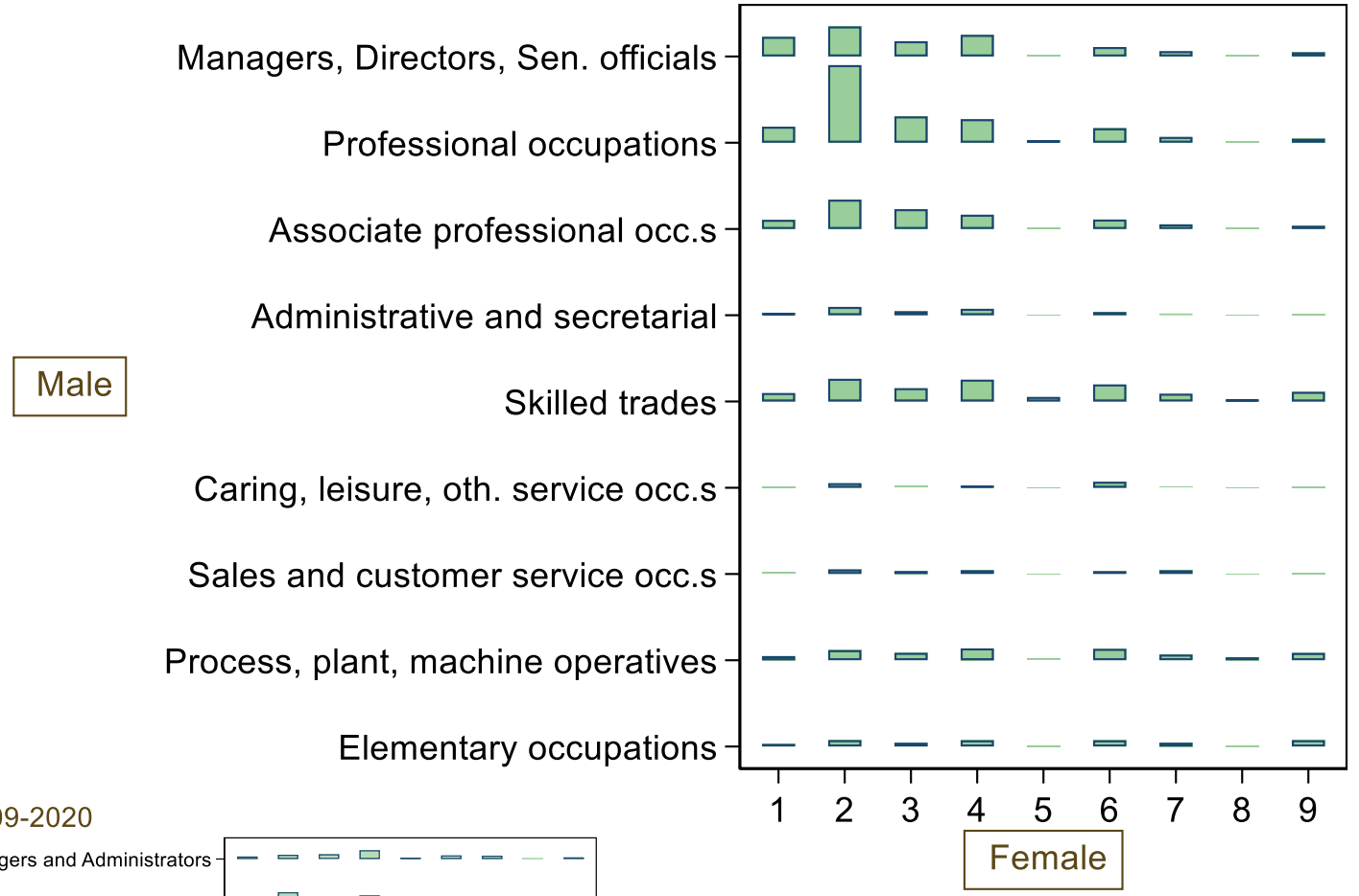
SOC 2010, 2010-20



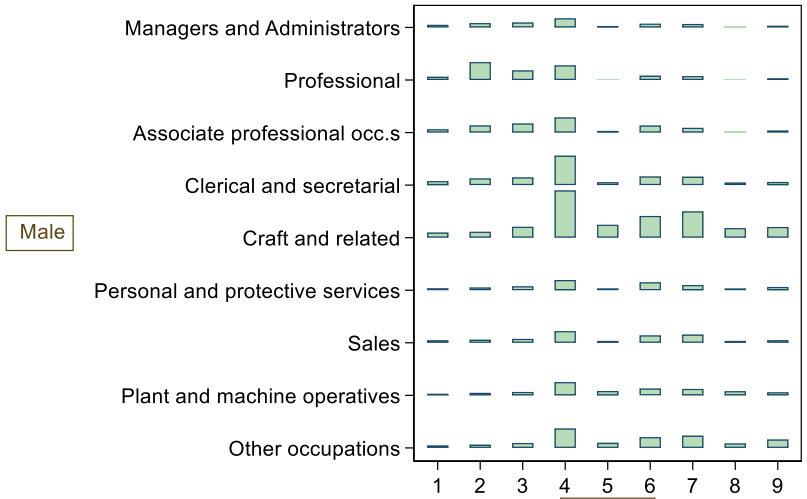
SOC 2000, 2009-20



SOC 2020, 2021-23



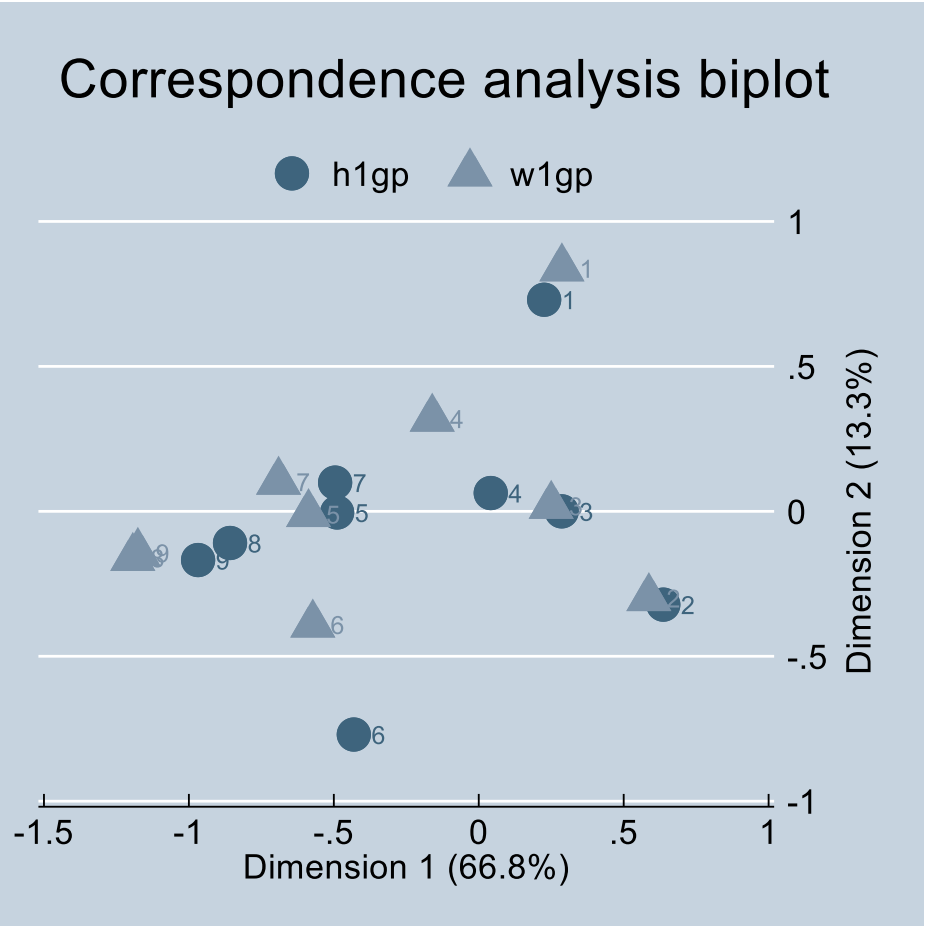
SOC 90, 2009-2020



Occupational coding frames' transformations...

Working towards a CAMSIS scale for SOC2020

- **www.camsis.stir.ac.uk/downloads/gb/gb_soc2020_v1.dta**
- **www.camsis.stir.ac.uk/downloads/gb/gb_soc2020_prep1.dta**



Quarterly Labour Force Surveys with SOC2020 since spring 2021

41261 records of both-working heterosexual couples (most recent unique occ pair, ~ 35k different couples)

'prep1' version an intermediate output based on standard methods

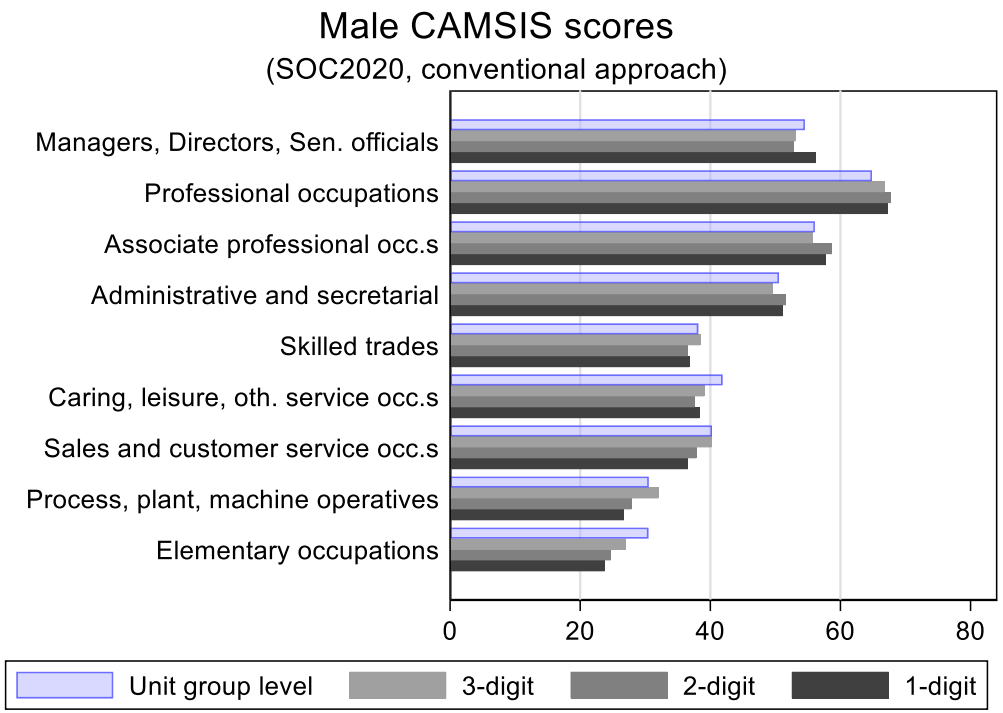
'v1' version uses additional bespoke adjustments (discussed below)

The datasets also feature NS-SEC (occupational level only), and ISEI, SIOPS and ICAM linked via ISCO08-88

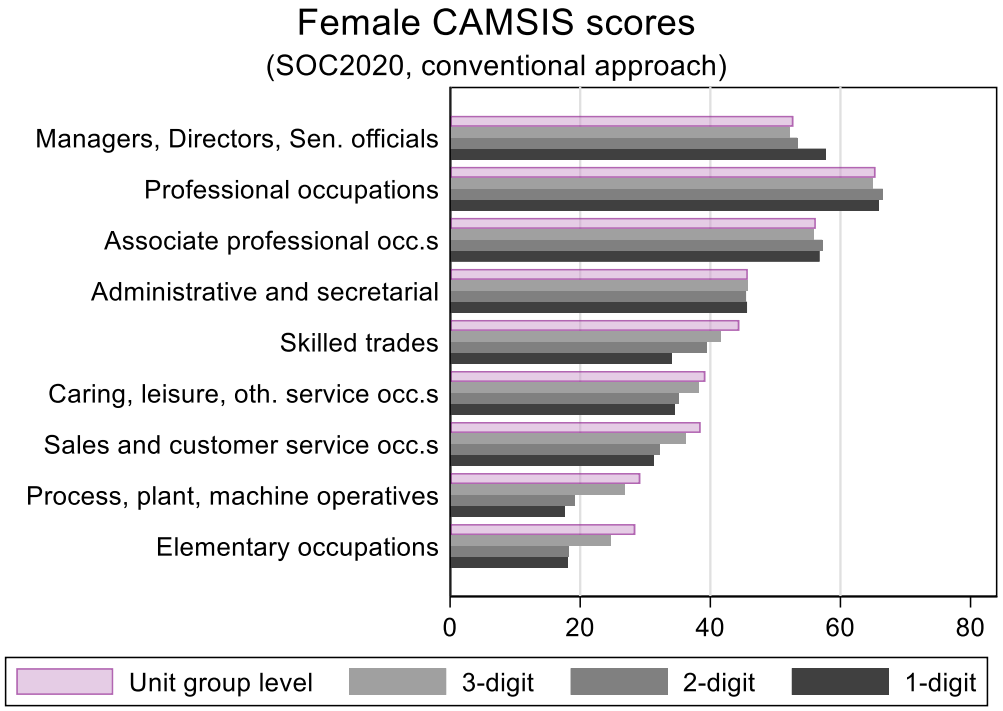
```
ca h1gp w1gp [fw=freq]
predict hscores, rowscore(1)
sum hscores [fw=freq]
replace hscores = ((hscores - r(mean)) / r(sd))*15 + 50
```

2) A version using standard routines

- **www.camsis.stir.ac.uk/downloads/gb/gb_soc2020_prep1.dta**

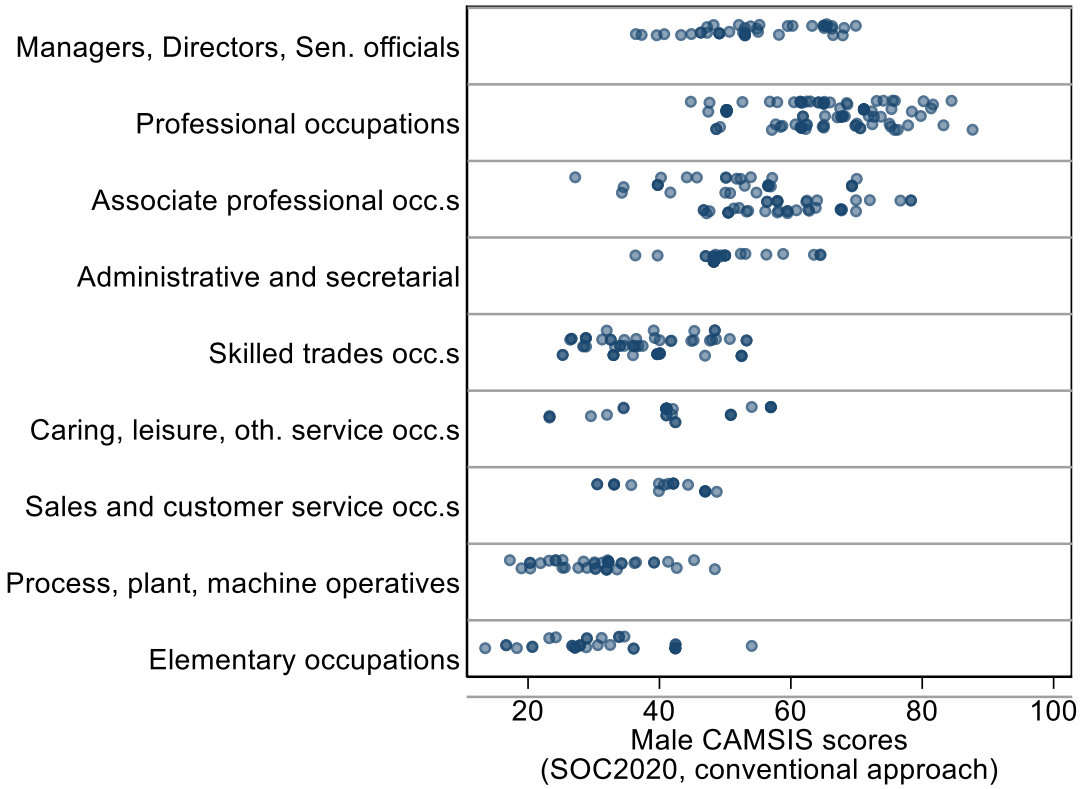


[This version based on standard CA-scores excluding unit group diagonals plus 'pseudo-diagonals' for farming, medics, catering and journalism]

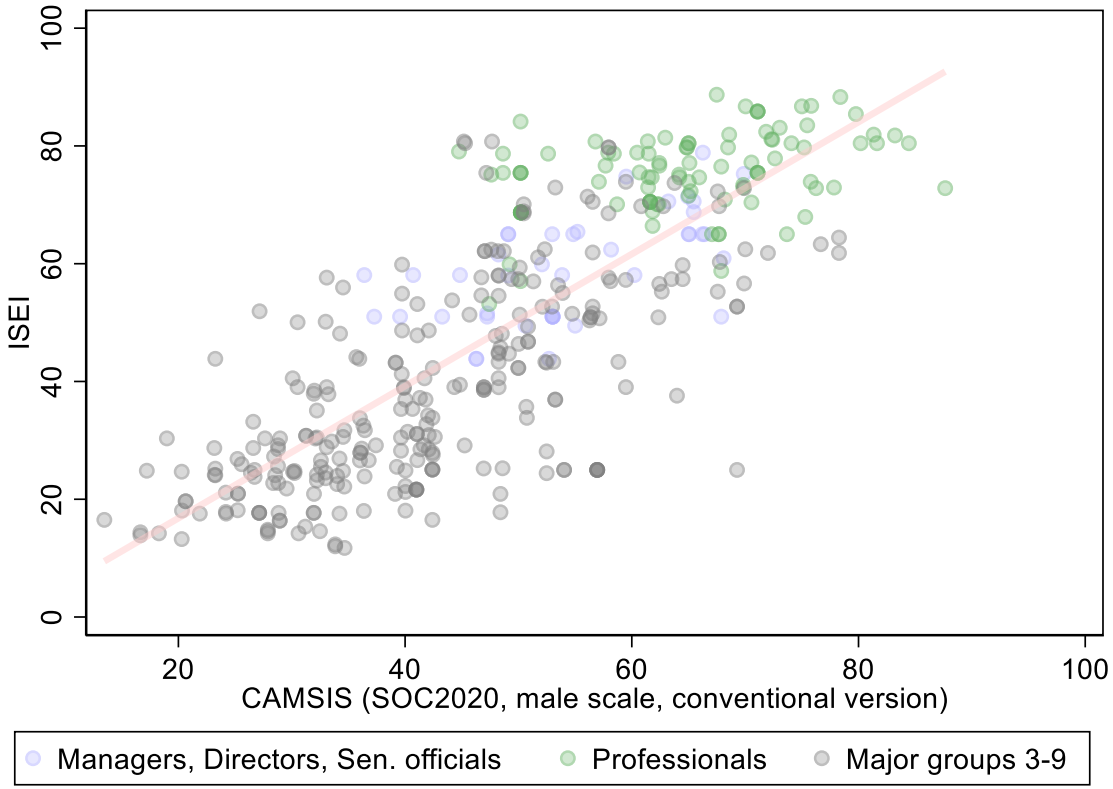


2) A version using standard routines

- **www.camsis.stir.ac.uk/downloads/gb/gb_soc2020_prep1.dta**



- *Correlations to ISEI/ICAM are a bit low (0.85/0.85)*
- *On inspection there are managerial jobs with very low scores and premium to major group 2 is pronounced*

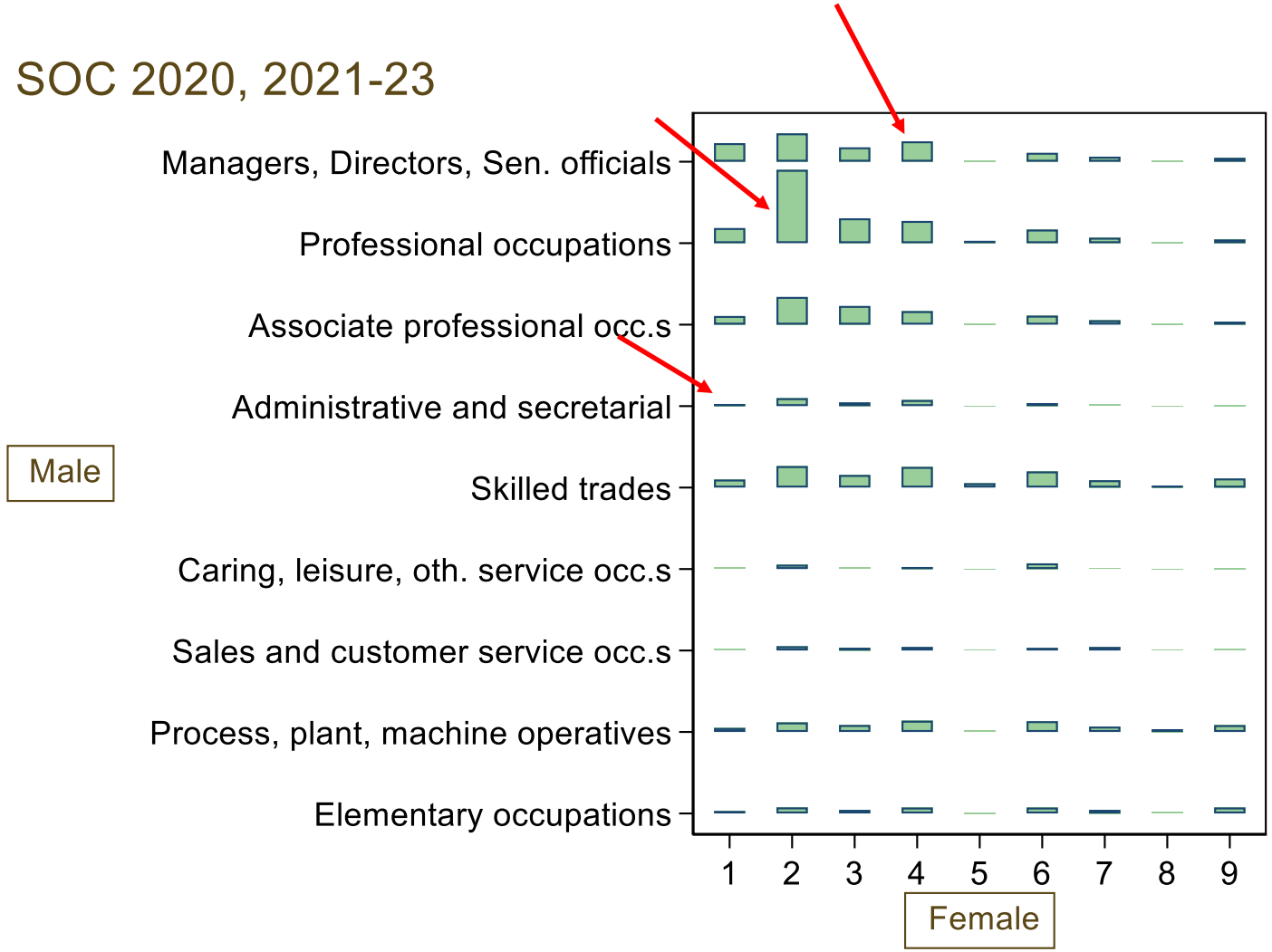


2) A version using standard routines

- **www.camsis.stir.ac.uk/downloads/gb/gb_soc2020_prep1.dta**

- *On inspection, most unexpected patterns are 'genuine' (e.g. psychologists rarely marry outside professional circles, but logistics managers do)*
- ***Might a 'professional homogamy' skew have emerged, that makes homogamy an imperfect indicator of stratification..?***
  - *[h1gp=1 + w1gp=4] an apparent PSD*
  - *Is [h1gp=2 + w1gp=2] a (university-driven) PSD?*
  - *Career/friendship/intergenerational data may be more suitable?*

➢ *...Try out options for adjusting methods to counteract any 'professional homogamy' skew..*



3) A version incorporating additional refinements

- **www.camsis.stir.ac.uk/downloads/gb/gb_soc2020_v1.dta**

- **Uses diagonals and pseudo-diagonals, but with more specification than previously common**
  - *Blanks out diagonals only in major groups 1 & 2, all H 1 to W 4 major groups, and psds for farmers, medics, catering, journalism, police, arts & housekeeping*
- **Uses a 'smoothing' strategy using random effects residuals to 'shrink' small group estimates towards the mean:**

```
ca hocc wocc if psd4==0
predict hscores, rowscore(1)
predict wscores, colscore(1)
mixed wscores if psd4b==0  ||hocc:,
predict hscores2, reffects
mixed hscores if psd4b==0 ||wocc:,
predict wscores2, reffects
```

- I also experimented with a weighting approach in which recommended scores were weighted averages of derived scores at different levels of occ. detail, but after exploration this didn't lead to compelling improvements in measures

	soc2020	mcamsis	fcamsis	isei	siops	icam
173.	1251. Property, housing and estate managers	64.08	63.21	62.39	49	59.95
174.	1252. Garage managers and proprietors	40.29	39.68	51.01	47	56.18
175.	1253. Hairdressing and beauty salon managers and proprietors	56.68	46.84	51.01	47	56.18
176.	1254. Waste disposal and environmental services managers	47.44	50.5	51.01	47	56.18
177.	1255. Managers and directors in the creative industries	73.39	79.12	65.01	75	60.13
178.	1256. Betting shop and gambling establishment managers	56.68	47.85	51.01	47	56.18
179.	1257. Hire services managers and proprietors	47.23	47.85	51.01	47	56.18
180.	1258. Directors in consultancy services	77.41	81.62	51.01	47	56.18
181.	1259. Managers and proprietors in other services n.e.c.	56.68	47.85	51.01	47	56.18
182.	2111. Chemical scientists	70.54	71.42	83.5	69	80.22
183.	2112. Biological scientists	69.74	71.42	80.46	62.66	68.98
184.	2113. Biochemists and biomedical scientists	77.52	71.42	80.46	62.66	68.98
185.	2114. Physical scientists	76.9	83.2	86.81	67	80.22
186.	2115. Social and humanities scientists	73	74.43	83.09	68.51	76.83
187.	2119. Natural and social science professionals n.e.c.	74.88	71.42	80.46	62.66	68.98
188.	2121. Civil engineers	60.38	59.75	81.4	70	73
189.	2122. Mechanical engineers	60.03	59.75	77.1	66	73
190.	2123. Electrical engineers	58.83	59.75	80.78	65	73
191.	2124. Electronics engineers	55.29	59.75	80.75	65	73
192.	2125. Production and process engineers	44.6	59.75	79.05	54	73
193.	2126. Aerospace engineers	61.58	59.75	77.1	66	73
194.	2127. Engineering project managers and project engineers	59.74	59.75	78.69	55	73
195.	2129. Engineering professionals n.e.c.	52.24	59.75	78.69	55	73
196.	2131. IT project managers	55.56	56.34	78.86	61.15	67.11
197.	2132. IT managers	58.06	59.96	78.86	61.15	67.11
198.	2133. IT business analysts, architects and systems designers	61.69	59.44	74.66	51	75.39
199.	2134. Programmers and software development professionals	64.06	70.97	74.66	51	75.39
200.	2135. Cyber security professionals	62.08	67.14	75.13	51	75.39
201.	2136. IT quality and testing professionals	58.39	67.14	74.7	51	75.39

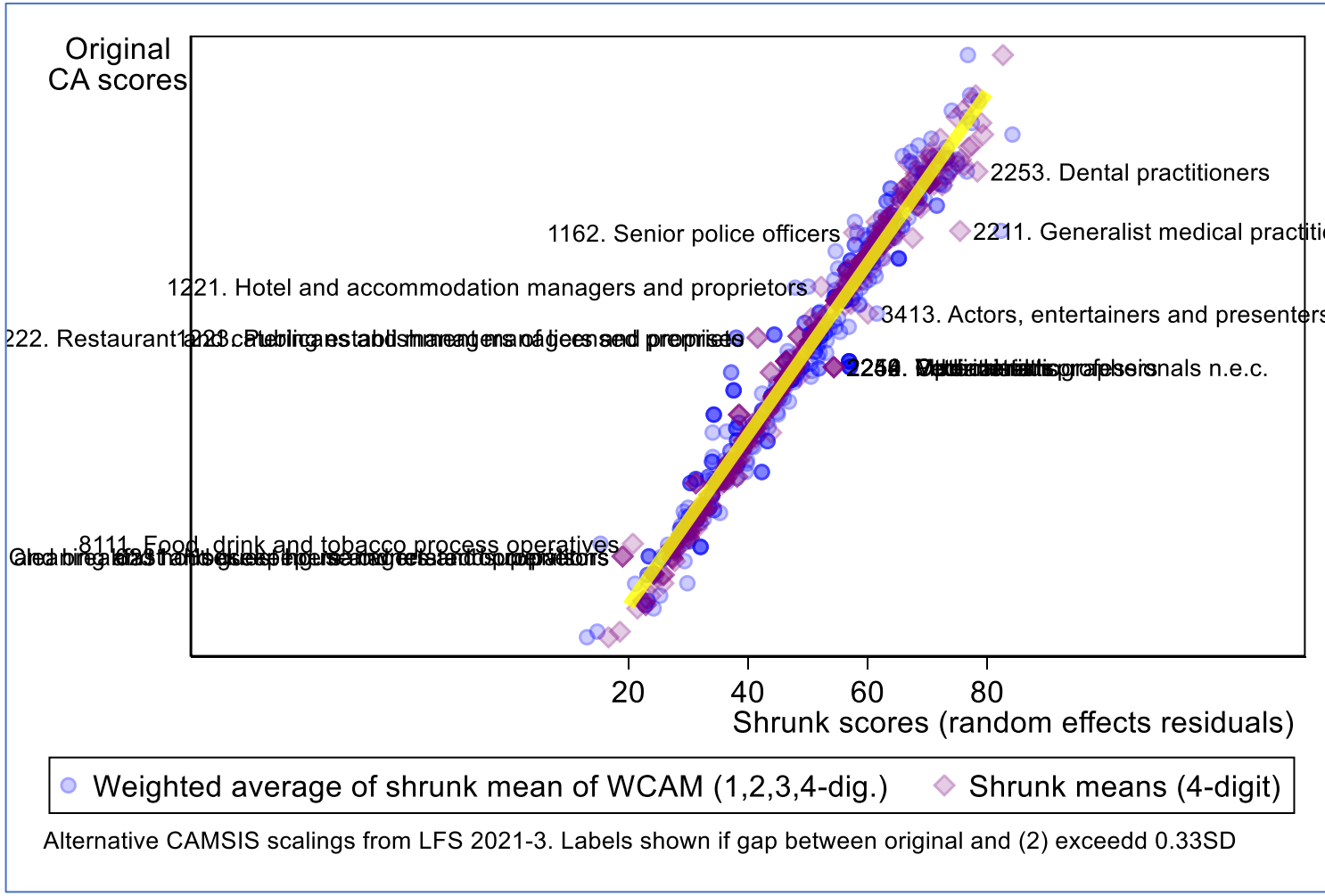
3) A version incorporating additional refinements

- **www.camsis.stir.ac.uk/downloads/gb/gb_soc2020_v1.dta**

- **Uses diagonals and pseudo-diagonals, but with more specification than previously common**
  Blanks out diagonals only in major groups 1 & 2, all H 1 to W 4 major groups, and psds for farmers, medics, catering, journalism, police, arts & housekeeping
- **Uses a 'smoothing' strategy using random effects residuals to 'shrink' small group estimates towards the mean:**

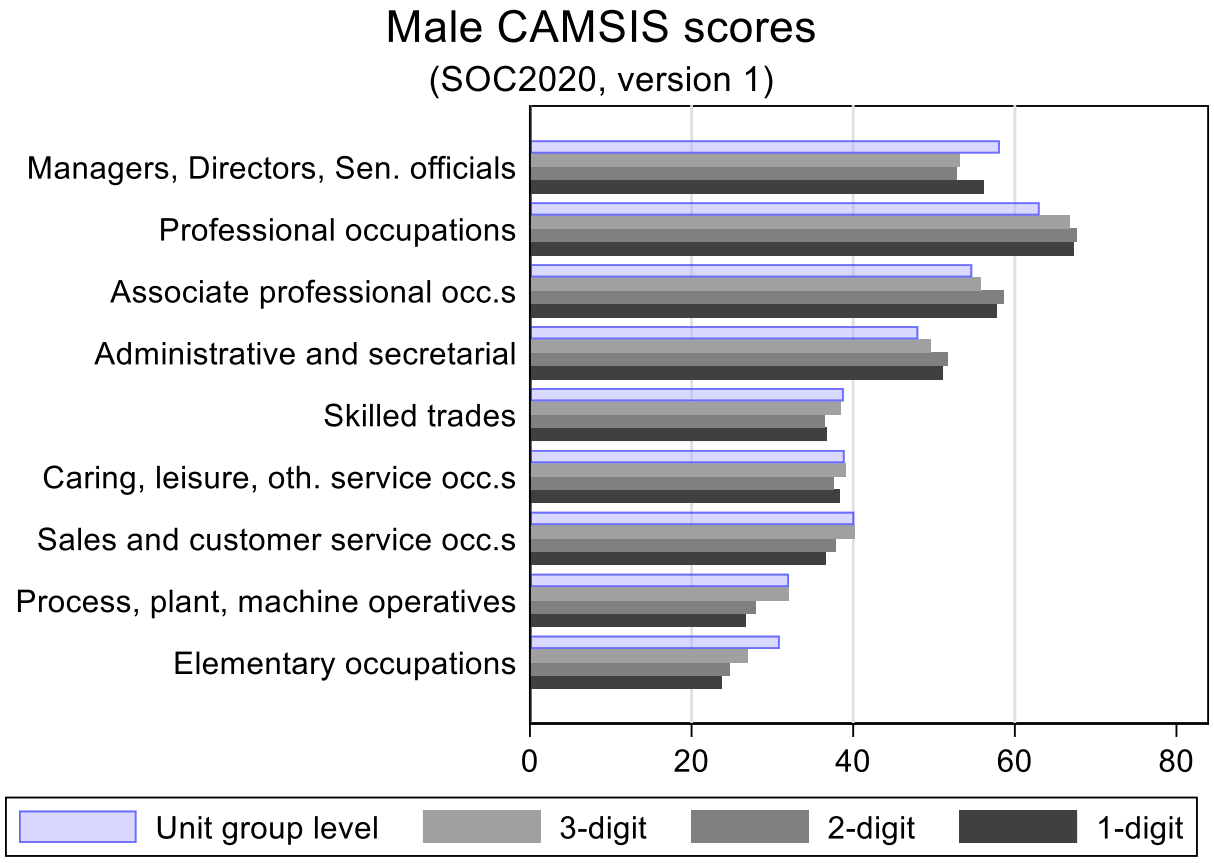
```
ca hocc wocc if psd4==0
predict hscores, rowscore(1)
predict wscores, colscore(1)
mixed wscores if psd4b==0  ||hocc:,
predict hscores2, reffects
mixed hscores if psd4b==0 ||wocc:,
predict wscores2, reffects
```

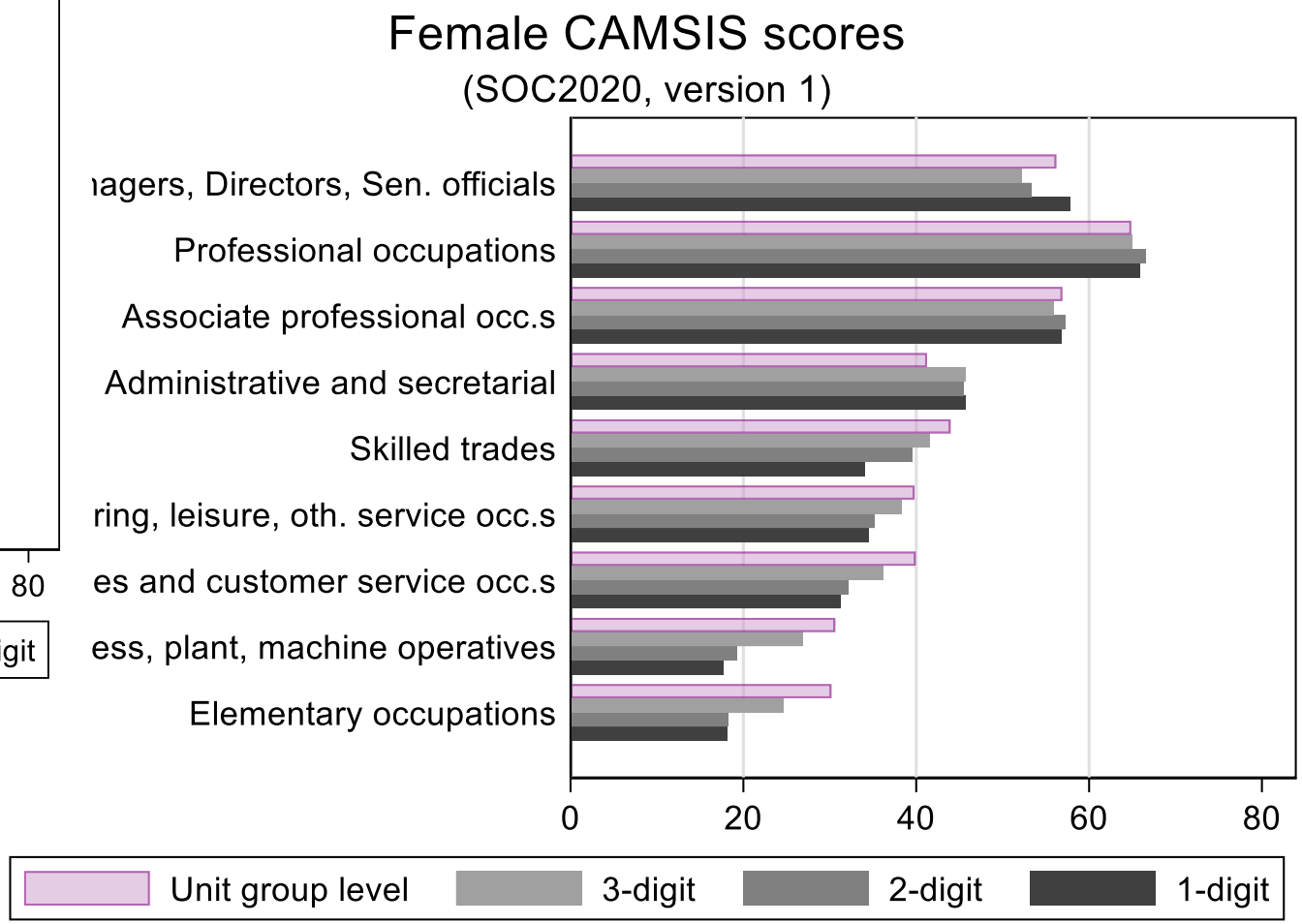
- I also experimented with a weighting approach in which recommended scores were weighted averages of derived scores at different levels of occ. detail, but after exploration this didn't lead to compelling improvements in measures



3) A version incorporating additional refinements

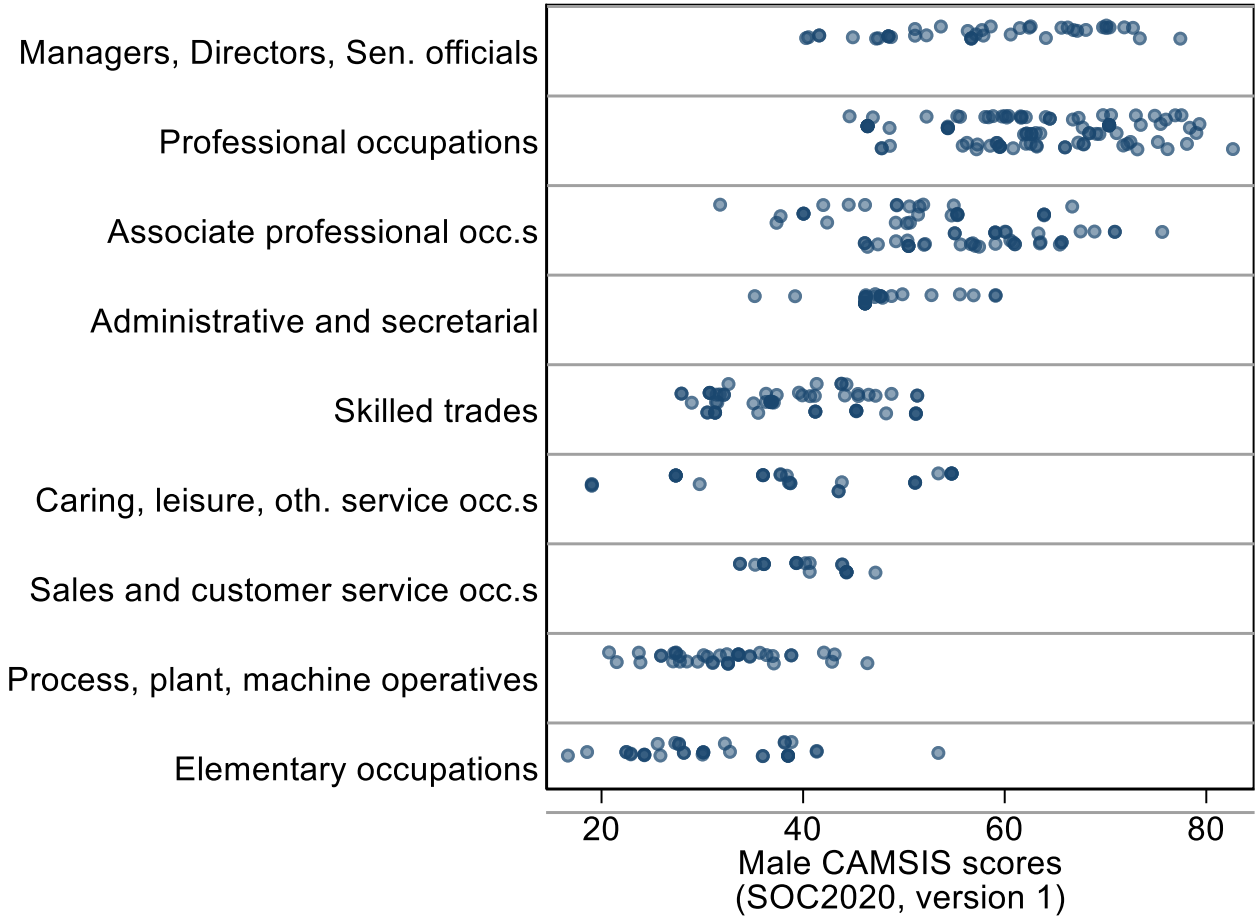
- **www.camsis.stir.ac.uk/downloads/gb/gb_soc2020_v1.dta**

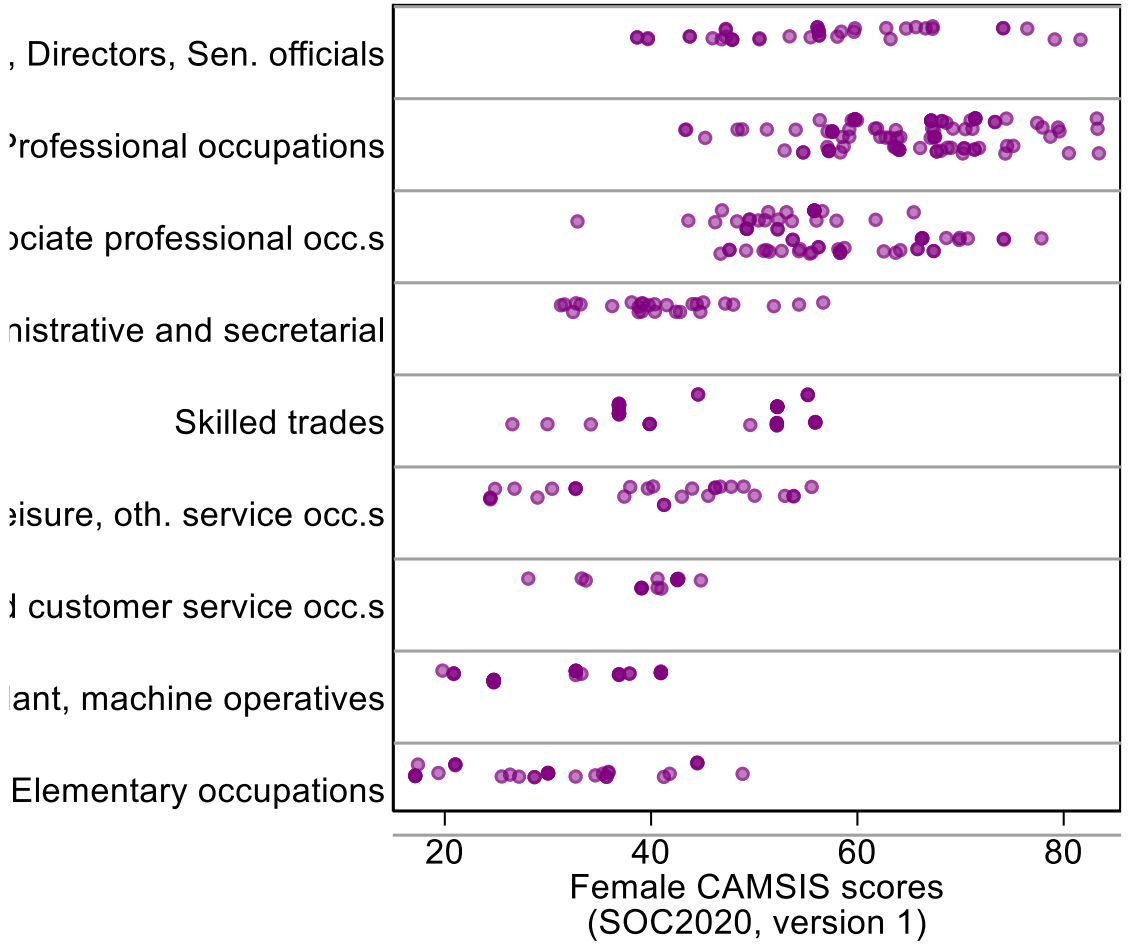




3) A version incorporating additional refinements

- **www.camsis.stir.ac.uk/downloads/gb/gb_soc2020_v1.dta**





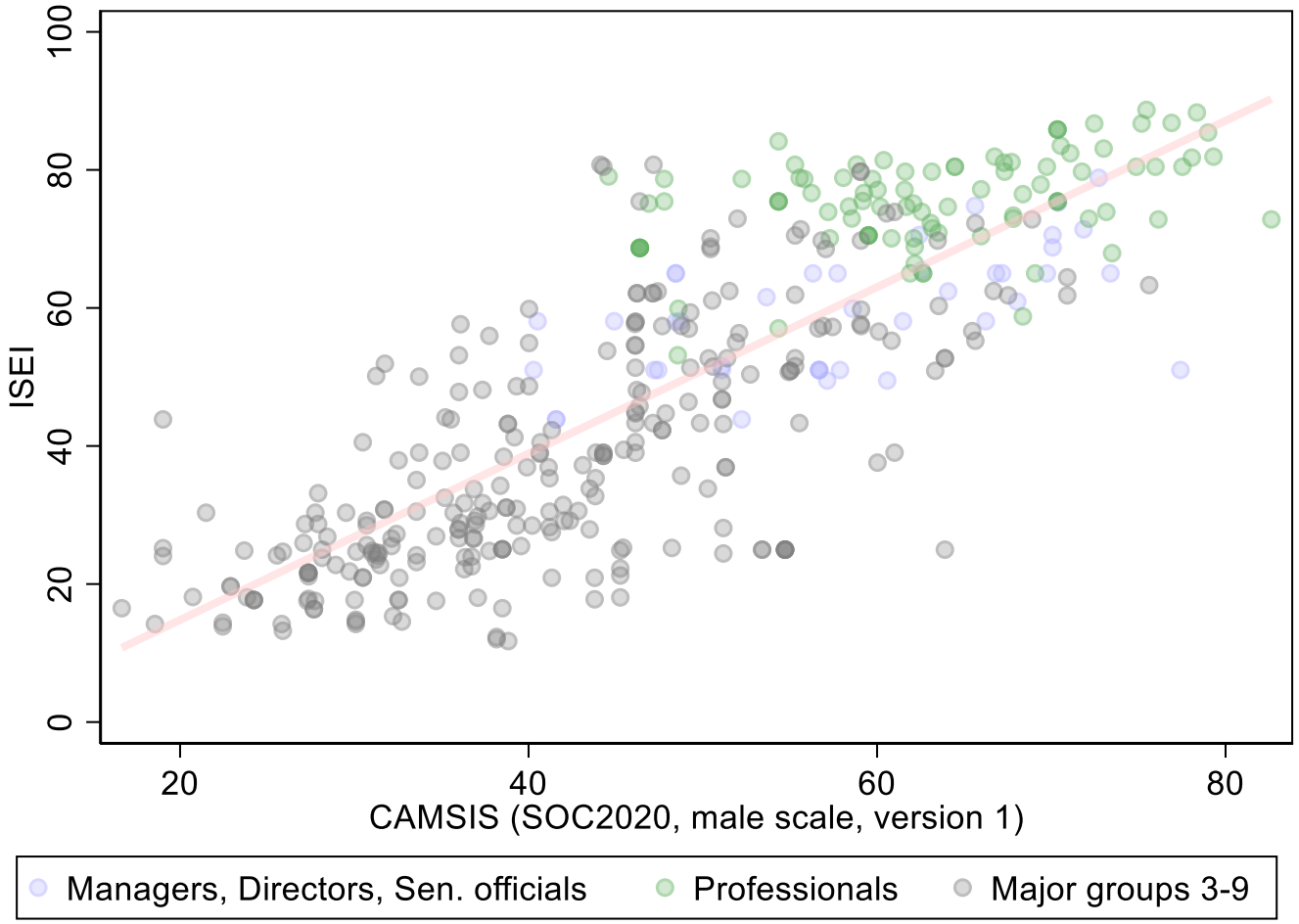
3) A version incorporating additional refinements

- **www.camsis.stir.ac.uk/downloads/gb/gb_soc2020_v1.dta**

- *Correlations to ISEI/ICAM are still a bit low (0.83/0.82)*
- *But on inspection, most divergences make sense and seem defensible, and the apparent skew of major groups 1 and 2 is reduced*

…'Additional refinements'

- Consistent with conventional CAMSIS methodology but seem destined to invite accusations of fixing the data...
- Could be labelled 'informative priors'...?



4) ‘Bivariate’ and ‘multivariate’ validity amongst the CAMSIS (gb_soc2020) measures

	MCAM	FCAM	MCAM_P	FCAM_P	ICAM	ISEI	SIOPS	NS-SEC
MCAM		87	98	85	80	81	75	80
FCAM	91		83	98	76	79	73	75
MCAM_P	96	86		81	80	80	73	78
FCAM_P	89	98	85		76	78	71	73
ICAM	83	82	84	82		92	82	84
ISEI	84	84	82	84	92		86	86
SIOPS	81	82	77	79	82	85		79
NS-SEC	84	84	81	81	86	89	81	

Top pane: At level of occupational unit groups (N-412); lower pane: At level of LFS microdata (AJ2023, N=24k adults aged 20-80)

Values are 100*Pearson’s R except for NS-SEC which shows 100*sqrt(R2) from regression predicting the linear term (NS-SEC 7-category version)

‘mcam’ and ‘fcam’ are the recommended versions, ‘_p’ variants are the preliminary versions without refinements

- *‘Bivariate validity’ would just be demonstrated by a variable correlating (in isolation) as expected with other things. It all but follows automatically from the high correlations between measures*

4) ‘Bivariate’ and ‘multivariate’ validity amongst the CAMSIS (gb_soc2020) measures

	MCAM	FCAM	MCAM_P	FCAM_P	ICAM	ISEI	SIOPS	NS-SEC
Gender	5	2	12	1	9	2	1	15
Age + Age2	11	10	11	10	12	13	11	8
Ethnicity	4	4	3	4	4	4	3	4
Social housing	27	26	25	27	25	27	25	27
Health prob.s	7	7	6	7	6	7	7	8
Disability	7	8	6	7	6	8	7	9
Log hrly. pay	46	49	42	47	45	50	45	53
Educ. Quals	50	50	51	50	49	51	47	26
Educ. lvg. age	12	12	13	12	12	11	10	7

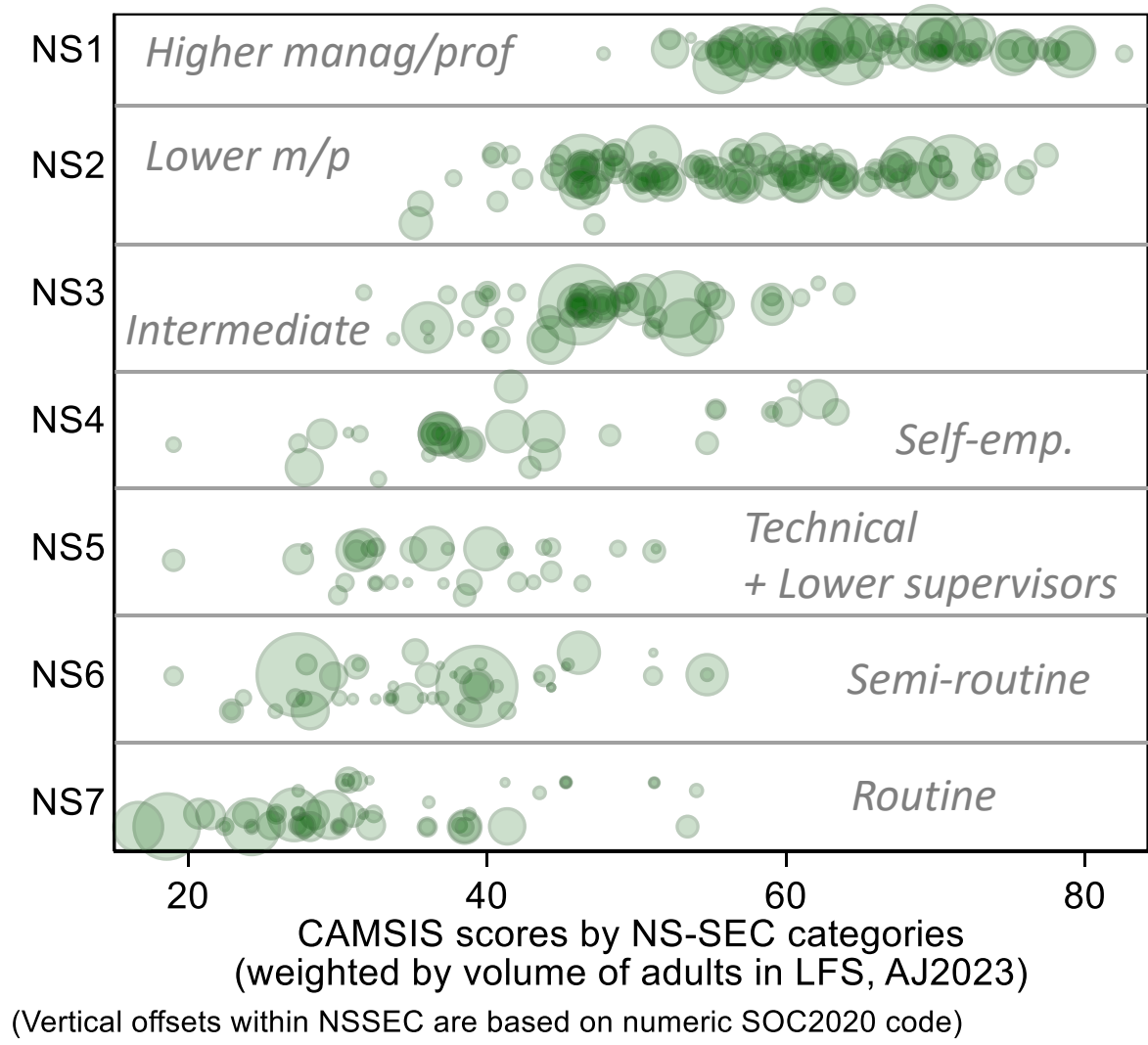
- *These results are comparable to previous studies which find much the same correlations between all stratification measures and other things.*

Values are 100*sqrt(R2) from bivariate models applied to LFS respondents (AJ2023, N=24k adults aged 20-80)

Models for gender, age, ethnicity and education predict the stratification measure, using linear regression for all except multinomial regression for NS-SEC. Logit models predict for social housing & disability as function of stratification measure. Linear regressions predict log hourly pay, education leaving age and health problems scale as function of stratification measure.

4) ‘Bivariate’ and ‘multivariate’ validity amongst the CAMSIS (gb_soc2020) measures

	NS1	NS2	NS3	NS4	NS5	NS6	NS7
	*(100*correlation between MCAMSIS and measure)*						
Social housing	6	8	6	13	12	9	10
Health prob.s	0	2	3	0	1	1	1
Disability	1	1	4	1	3	1	3
Log hrly. pay	9	13	0	30	32	10	4
Educ. Quals	27	32	9	41	10	9	19
Educ. lvg. age	16	13	2	20	4	7	10
N for (1)	5845	6732	3818	1873	1179	2395	2590



- *Correlations observed between MCAMSIS and measure within NS-SEC categories (i.e. heterogeneity within NS-SEC)*

4)…'multivariate validity' amongst …measures

- *The basic idea is that measures work fine if there is one and only one of them….*

Variable	mcam_1	isei_1	nssec_1	mcam_2	isei_2	nssec_2
z_mcamsis	-.0557***			-.00919		
z_isei		-.0584***			-.024	
ns_sec						
1. Higher..			(base)			-.0512
2. Lower ..			.0651***			(base)
3. Interm..			.119***			.0572
4. Small ..			.117***			-.0331
5. Lower ..			.0578*			-.187*
6. Semi-r..			.185***			.0295
7. Routin..			.196***			.0344
fem				.0962***	.0928***	.085***
age				.00785***	.00784***	.00718***
eth_black				-.118***	-.11***	-.116***
eth_asian				-.0638*	-.0588*	-.0607*
eth_other				-.115***	-.109***	-.112***
z_mcam_age				-.00098*		
z_isei_age					-.000644	
ns_sec#c.age						
1. Higher..						-.000039
ns_sec#c.age						
3. Interm..						-.000433
4. Small ..						.00164
5. Lower ..						.00453*
6. Semi-r..						.0018
7. Routin..						.00214

Bivariately valid in models that predict 'health problems' scale (LFS AJ23, n=24k)

4)…'multivariate validity' amongst …measures

- *…but it all goes to pot if they are purported to disentangle each from each other…*

Multivariately invalid
(as representations of stratification association net of stratification)

Variable	mcam_3	isei_3	nssec_3	mcam_4	mcam_5	mcam_6
z_mcamsis	-.0549***			-.0333***	-.0201*	-.0212*
fem	.0971***	.0933***	.0809***	.0957***	.0875***	.0833***
age	.00796***	.00792***		.00793***	.00796***	
eth_black	-.118***	-.11***	-.139***	-.115***	-.118***	-.142***
eth_asian	-.0637*	-.0586*	-.0968***	-.061*	-.0613*	-.0976***
eth_other	-.114***	-.109***	-.153***	-.111***	-.112***	-.153***
z_isei		-.0537***		-.0256**		-.00844
ns_sec						
1. Higher..			-.0479***		-.0417**	-.0317*
ns_sec						
3. Interm..			.0416*		.0248	.0209
4. Small ..			.0692**		.0251	.0363
5. Lower ..			.0167		-.00825	-.0245
6. Semi-r..			.117***		.0834***	.0723*
7. Routin..			.147***		.0955***	.0892**
_cons	-.675***	-.672***	-.323***	-.673***	-.685***	-.31***
N	24432	24432	24432	24432	24432	24432
ll	-29159	-29161	-29353	-29155	-29143	-29349
bic	58388	58393	58817	58391	58418	58830
r2	.0271	.0269	.0115	.0274	.0283	.0118

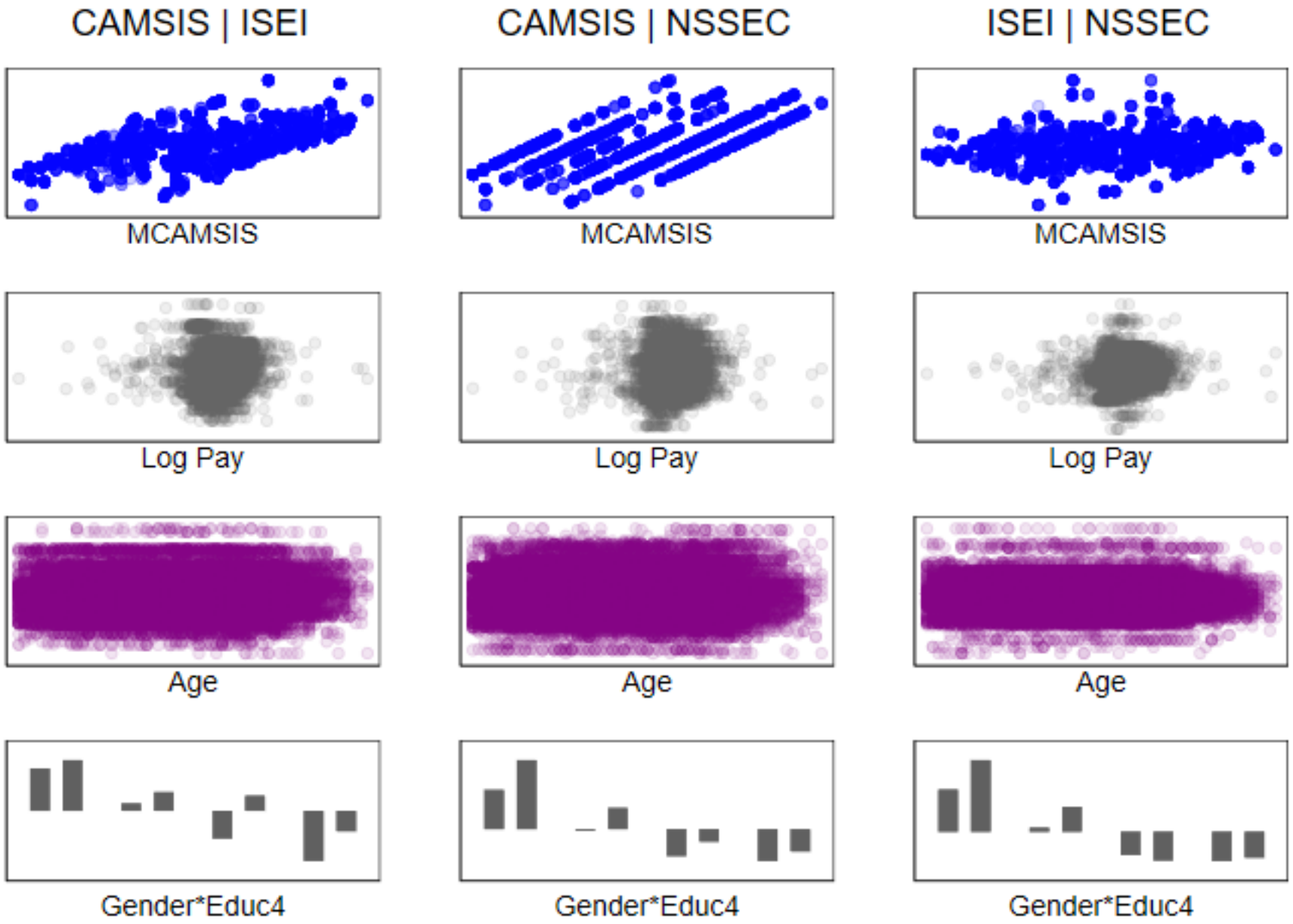
Bivariately valid in models that predict 'health problems' scale (LFS AJ23, n=24k)

4)…'multivariate validity' amongst …measures

- …In the regression framing, I think we can show multivariate invalidity by showing bivariate invalidity amongst residuals…

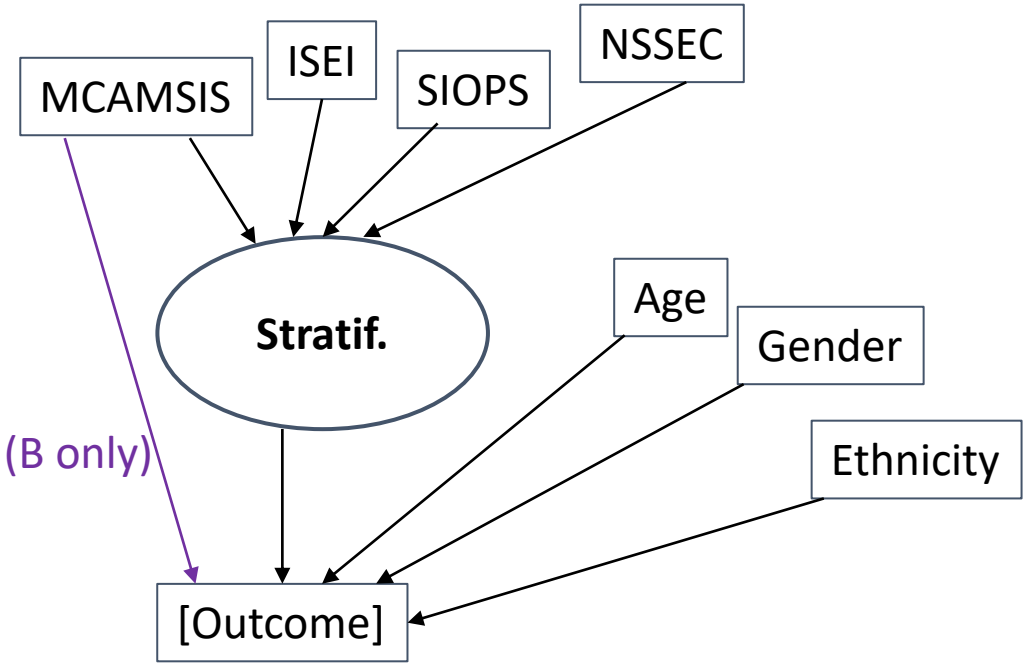
➢…such patterns imply residuals reflect **measurement errors** not meaningful dimensions (?)

Results below for LFS microdata, vertical axis for residuals
(i.e. calculate residuals from Y | X, then look at those residuals against other things)



4)…'multivariate validity' amongst …measures

- In the SEM framing, can work towards model fit comparisons by comparing formulations which force indicators onto a single latent factor, versus others that allow them to be work separately



Model	Outcome	RMSEA	AIC	L1	MC_B
A	Health probs.	0.144	228615	-0.065*	-0.018
B	Health probs.	0.145	228616	-0.046*	
A	Log pay	0.144	49140	0.588*	
B	Log pay	0.145 (sic)	49137	0.685*	-0.092*
A	Educ4 (as scale)	0.144	225265	-0.591*	
B	Educ4 (as scale)	0.145 (sic)	225165	-0.411*	-0.173*

- ..these examples contrast a 'one-factor' and 'two-factor' solution. The one-factor model always seems more succinct, but the two-factor solution may be a better fit to the data.
- Interpretation: Figures consistent with view that different measures can mop up measurement error of each other, but without revealing fundamentally different stratification processes (?)

Summary: A CAMSIS scale for the UK's Standard Occupational Classification 2020 and reflection on its 'multivariate validity'

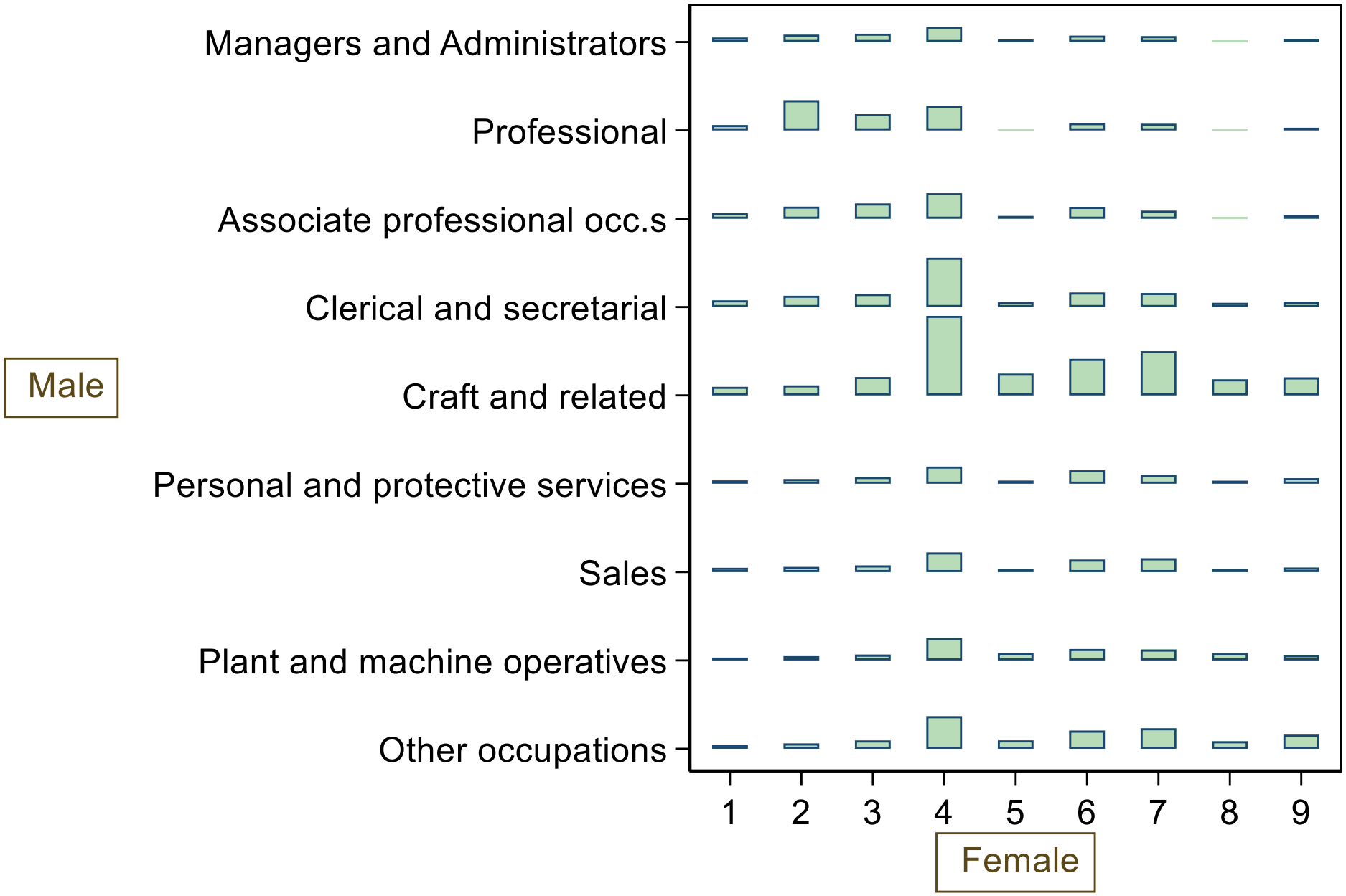
www.camsis.stir.ac.uk/downloads/gb/gb_soc2020_v1.dta

www.camsis.stir.ac.uk/downloads/gb/gb_soc2020_prep1.dta

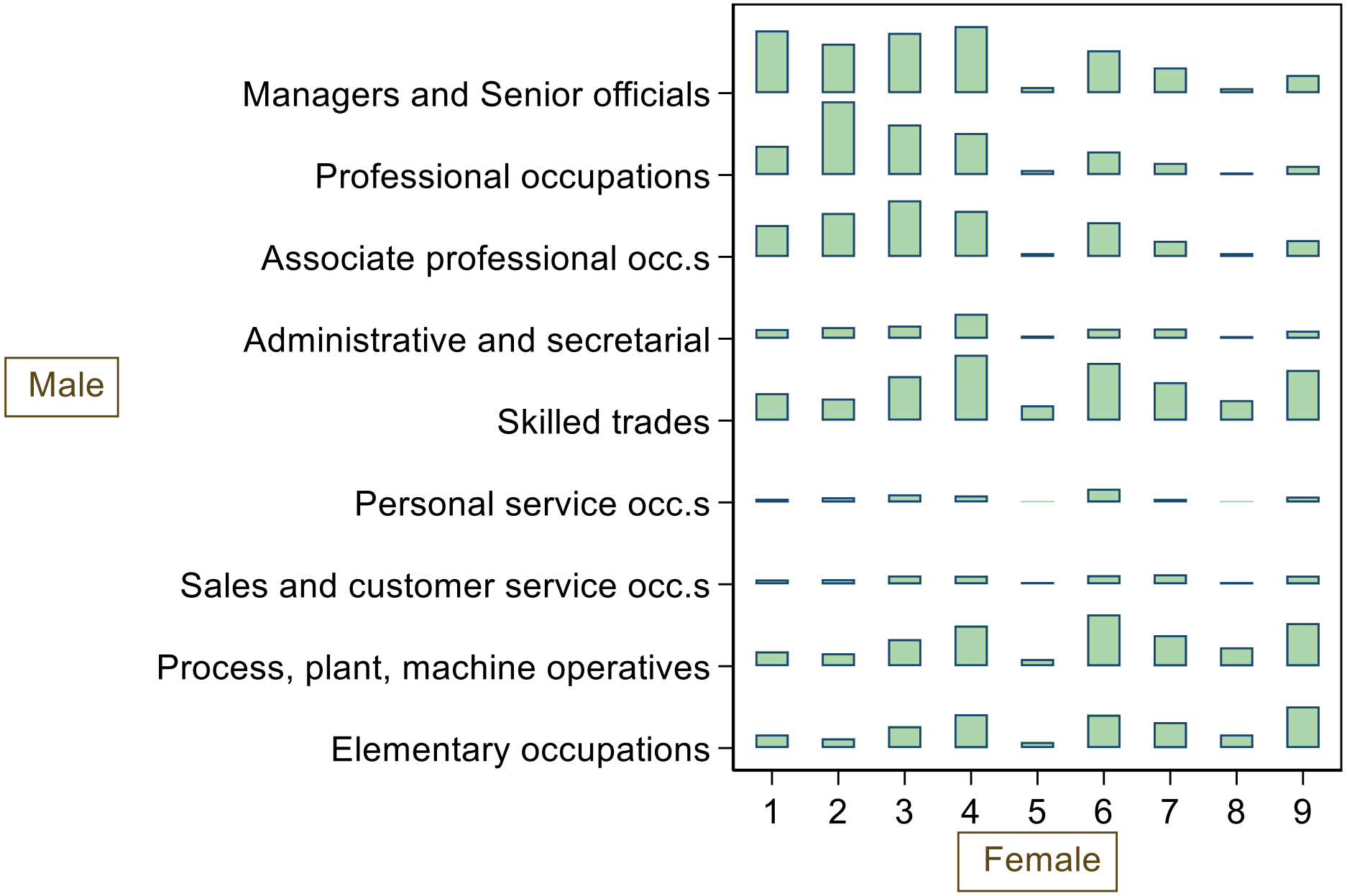
- A fairly compelling scale that is a parsimonious indicator of position within the structure of social stratification and its reproduction
  - Linked to education in particular
  - Professional v Managerial advantaged positions of methodological note
  - More datasets with SOC2020 units to support further comparisons (only LFS hitherto)
- Valid as a stratification indicator
- Invalid as a consistent measure of a different thing to stratification
- Not exactly the same as other stratification measures, but gaps between them aren't consistently interpretable (suggestive of measurement error)

Appendix - Some of the same images, but bigger…

SOC 90, 2009-2020



SOC 2000, 2009-20



SOC 2010, 2010-20