

LMIP pre-colloquium Workshop 2016

Sample selection in the WCGDS:

Analysing the impact for employment analyses

Nicola Branson

Aims

- Examine non-response in the WCGDS in detail in order to:
 - Document procedures used in the survey and their impact on outcomes to provide a resource for those using these data and for those planning the National GDS
 - Describe baseline differences in characteristics between those who responded versus those who did not respond to the survey and explore these in a multivariate framework
 - Propose methods to assess and account for non-response bias using national administrative databases and information on the type of contact details available

Graduate Destination Studies versus Household Survey Data 1

- GDS provide detailed information about graduate transitions to work that cannot easily be collected in household surveys.
- Graduates represent $< 0.3\%$ of population, therefore, nationally representative household study samples are not large enough to disaggregate by institution or study program.
- Only contain concurrent socioeconomic information. Makes it difficult to disentangle the impact of studying from preexisting characteristics.

Graduate Destination Studies versus Household Survey Data 2

- Panel studies such as the **Cape Area Panel Study** and the **National Income Dynamics Study** go a step further and provide more detail on the factors associated with who attends and who does not.
- Though continue to suffer from the same sample size issues.
- Therefore used to answer questions of a more aggregate nature e.g. who attends university versus TVET versus nothing.

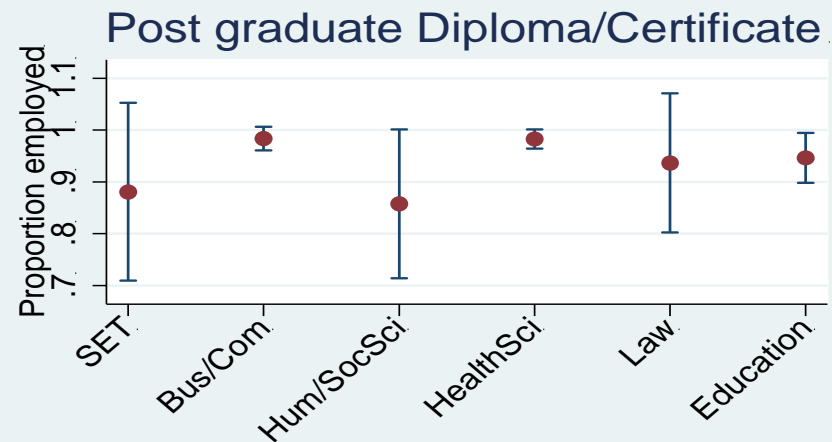
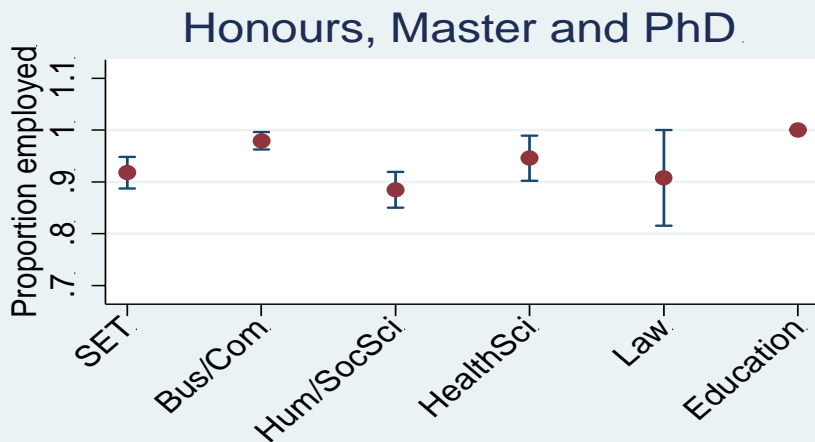
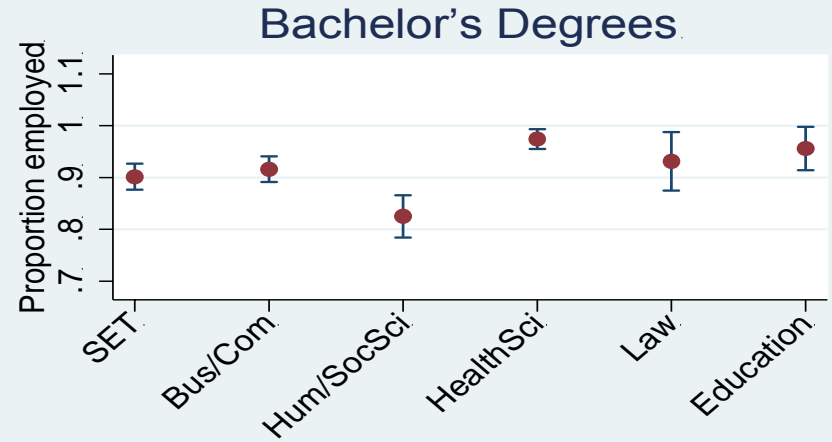
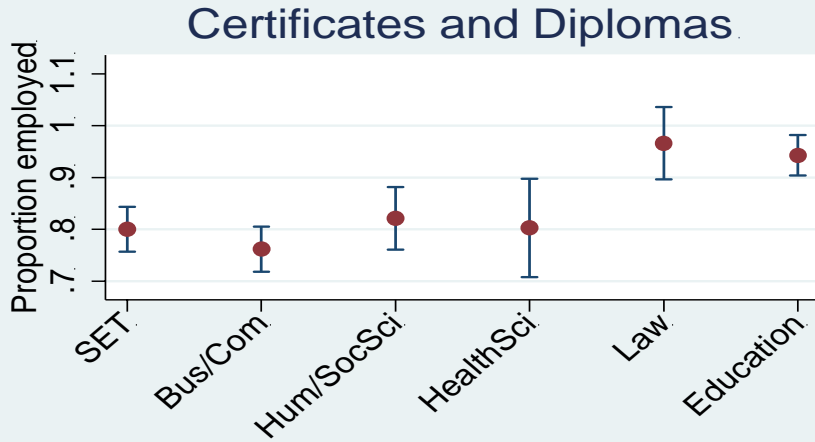
Graduate Destination Studies versus Household Survey Data 3

- By focusing on the graduate population, graduate destination studies circumvent the sample size issues.
- The focus on graduates allows a disaggregation by institution and field of study and allows investigation into the match between labour force participation and labour shortage areas.
- Enable unpacking of puzzles within the higher education sector that are often unexplored due to data limitations.

Western Cape Graduate Destination Study

- Attempted to contact all 2010 graduates from the four WC HE institutions in 2012, two years after graduating.
- Run by Cape Higher Education Consortium (CHEC).
- Voluntary responses – 5560 graduates responded.
- Include questions on:
 - the labour market and further studying trajectories
 - perceived value of their qualification once working,
 - relevance in the workplace and how much the qualification prepared the graduate for work
 - satisfaction with work obtained.

Employment rates of graduates by qualification type and area of study



weighted.

GDS's achilles heel – low response rates

campus	mean	N
CPUT	0.218	7441
UCT	0.219	6165
US	0.216	7380
UWC	0.267	3724
Total	0.225	24710

- WCGDS response rates

Non-random non response

- Non response is always a concern in that it reduces sample size and therefore the power of a survey to demonstrate relationships of significant interest.
- Main concern for a survey of this size is that those who respond are different in important ways to those who do not respond.
- Given the large sampling frames generally available to utilize in the design of a GDS – all graduates – this concern is doubly problematic as the **realised samples are usually large enough to get precisely measured estimates even if these estimates are in fact wrong**

Thinking about response patterns: Survey process details

- Institutional differences in the completeness of contact details.
- HEMIS data was from 2010, contact details came from the most up-to-date records on the institutional software.
- Contact details from the National Study Financial Aid Scheme (NSFAS) only sought for student at CPUT and UWC.
- (Almost) All graduates sent an email with a cover letter signed by institution's dean
- A non-random telephone follow up was used to increase the response rates for graduates predominantly from CPUT and UWC.
 - No information is available on who the call center attempted to call, only on the mode of interview (email or phone) for those who did respond

Differences in baseline characteristics

- UCT characteristics between responders and non-responders are more balanced than in other institutions, CPUT is the least balanced.
- **Higher share of females** in the responder group at each institution, but especially at CPUT and UWC.
- Responders are significantly **more likely to be African** (at all except UCT) especially at CPUT where the difference is 9% points.
- The targeting of NSFAS students at CPUT and UWC is evident in the data, with the **share of NSFAS bursary holders much higher** in the responder group.
- **Higher share of other bursary holders** at all institutions. This could be a result of better contact details or signal a stronger connection to the institution.

Differences in exit characteristics

- Vary across institution.
- Higher share of Masters students among UCT and US responders.
- Other differences indicate that **responders are on average more advanced students at UCT, US and UWC but less highly qualifying students at CPUT.**
- Few program differences are apparent although **SET students appear overrepresented** in the responder group at all institutions.

Existing Weights in the WCGDS

- Responses were linked to the HEMIS database of all 2010 graduates.
- Successfully contacted graduates were statistically weighted to the actual socio-demographic profile of the 2010 cohort on:
 - **gender, population group, qualification type and institution**
- Using the weights reduces differences on the variables included in the reweighting but other differences remain.

What the weights do not do

- To extent that these (and no other observed or unobserved) characteristics determine the outcome analyzed, the data are representative
- Account for the fact that those who respond and those who do not respond, could be systematically different in other ways
- Other characteristics (e.g. migration, employment, motivation) can affect success in the labour market or other outcomes of interest *and* the probably of completing the survey

Contact details by response outcome

	Non-Responder	Responder	Responder weighted
Contact details available:			
Cell	0.66	0.88***	0.87***
Landline	0.57	0.55	0.56
Email	0.71	0.77***	0.76***
NSFAS cell	0.13	0.17***	0.15***
NSFAS email	0.12	0.17***	0.15***
Email type:			
UCT	0.04	0.03***	0.03***
CPUT	0.01	0.01	0.01
UWC	0.11	0.12**	0.1**
SUN	0.17	0.14***	0.14***
institutional	0.34	0.3***	0.29***
Sample size	19150	5560	5560

When is non-response a problem?

- When non-response is non-random and related to the variable of interest.

$$y_i = x_i' \beta_1 + \varepsilon_i \quad y_i \text{ observed only if } A_i^* < 0 \quad (1)$$

$$A_i^* = x_i' \beta_2 + z_i' \gamma + v_i \quad \text{selection equation} \quad (2)$$

- Non-response is a problem if ε_i and v_i are correlated
- In this case β_1 estimates will be inconsistent
- From the model it is clear that an evaluation of non-random responses is model-specific

Assessing the impact of non response for employment analyses

- Bias in observables:
 - Non-response probit
 - BGLW test
- Unobservables
 - Using external data to check the composition
 - Heckman selection model using institutional email as an instrument

Testing whether response is random: Selection Probit

- Construct a dependent variable takes the value one for individuals who drop out of the sample after the first wave (attrit) and zero otherwise.
- Run a probit regression
- Explanatory variables are baseline(2010) values for:
 - all variables that are believed to affect the outcome variable of interest (employment),
 - plus any available variables which characterise the interview process.
- Chi-square test for joint significance of characteristics included
- Tests are model specific and need to be repeated for each outcome variable of interest.

Selection Probit results

- Type of contact information available strongly predictive of response
- Quite a lot of variation across institutions
- Whites less likely to respond at all institutions
- CESM strongly predictive of response
 - All CESMs are less likely to respond than SET
- Type of qualification predictive of response rates at UCT and US, but less so for other institutions
 - The higher the qualification the more you are likely to respond at UCT and US

Selection probit results

- The Chi-square test statistics show that response is non-random for all institutions.
- One way to adjust for this non random non-response would be to construct a weight equal to the inverse probability of not responding.
- However, the R--squares, especially for UCT and US, are small therefore suggest that the impact of reweighting exercise is going to be small.
- Also this approach would not preclude there being other unobserved or unmeasured characteristics that could bias the results.

Testing for sample selection using BGLW pooling-test

- Another common test for sample selection is the pooling test due to Beckett, Gould, Lillard and Welch (1988)
- The BGLW test involves regressing an outcome variable available for all graduates on
 - household and community variables,
 - An indicator that the graduate did not respond,
 - The non response indicator interacted with the other explanatory variables.

Testing whether non-response is random BGLW-test

- Don't have employment information
- But For the UCT subsample we have information at the individual level of who was enrolled at UCT in 2012 from the UCT administrative database
- We run linear probability models for the probability of studying in 2012 on the LHS
- On the RHS we have:
 - “usual” explanatory variables,
 - non response indicator
 - and all the variables interacted with non response indicator

UCT sub analysis: Testing whether response is random BGLW-test

- The non response coefficient is -0.13 (s.e. 0.08) which while not significant confirms that those who did not respond are less likely to be studying in 2012.
- In addition, two of the interaction terms – age and other bursary – are statistically significant
 - Relationships weaker for the non-responder group
- The F-stat of 2.19 and the p-value of 0.000 make us reject the null hypothesis that response is random, at least for the probability of studying in 2012

Complementary administrative data

	All 2010 Graduates			WCGDS Responders		
	Number with ID numbers	In 2012 HEMIS database		In 2012 HEMIS database	Studying at university according to Q4_1	
	#	#	%	#	#	%
CPUT	6894	1,450	21%	426	430	27%
UCT	5102	1,143	22%	300	379	28%
US	7380	1,710	23%	492	533	34%
Total	19376	4,303		1218	1342	

Taking stock

- We have shown that the composition of the responder sample differs from the non-responder sample
- The UCT specific analysis shows that in addition to the compositional difference, the relationship between the determinants of studying and study probability differs for those who respond versus do not respond
- Suggests that the characteristics of WCGDS responders who are studying in 2012 are not representative of the full studying population.
- In addition, the covariates included only explain 10% of the probability of studying suggesting that there are other unobserved characteristics that explain studying probability, which in turn could also differ by responder status.

What can we do

- We have evidence that response in the WCGDS is not random.
- How should we deal with this problem?
- Heckman selection model – if you have a believable instrument.
- Reweighting – if you believe the observed baseline characteristics determine the selection.

Selection corrected employment models

- This approach requires an exclusion restriction, a factor z_i' that is correlated with response but not correlated with ε_i in equation (1)
- Argue that institutional email will not be related to employment probability when the sample is restricted to the labour force except possibly through some of the characteristics included in the structural equation.
- Run a regression of the probability of employment for those in the labour force.

Selection corrected employment models

- Inverse mills ratios are negative in all except the CPUT model
- Not significant in any of the models
- Suggests that, conditional on the assumptions of the model, selection does not appear to be a significant problem for this model.
- Ignoring significance, the direction of the lambda coefficients suggests that UCT, US and UWC graduates who responded to the survey were less likely to be employed while CPUT graduates were more likely to be employed.

Conclusions

- Destination studies have a particular type of bias which is inherent to their design
- The characteristics of responders and non-responders differ in non-random ways and this is clear when looking at observable or measureable variables.
- The direction of the bias (on employment outcomes) that may result from these observable differences between responders and non-responders is difficult to identify in the WCGDS data.
- Re-weighting on a subset of observable predictors of non-response only accounts for part of the bias.
- Responders/non-responders differ in unobservable ways for which it is not possible to adjust with statistical weights.

Conclusions

- Fortunately there is an approach which has been used in the literature which will allow for some type of control for selection based on unobservable characteristics.
- In the WCGDS, the method suggests that selection bias does not appear to be a major concern for analyses of employment outcomes. We therefore have some confidence in the estimates of employment probability from these data.
- These findings only apply to employment outcomes and the approach described in this paper would need to be conducted again for other outcomes (e.g. job satisfaction, job matching, or earnings) of interest.

Recommendations

- Preparation and consistency sampling frame with comprehensive baseline information vital, especially contact details.
- Record information about the survey process at an individual level.
- Link/triangulate data and findings with other administration data to assess bias in key estimates.

Thank you

Some Readings

- **Baulch, B., & Quisumbing, A.** “Testing and Adjusting for Attrition in Household Panel Data”. *CPRC Toolkit Note*, 2011.
- **Magruder, J., & Nattrass, N.** “Exploring Attrition Bias: The Case of the Khayelitsha Panel Study (2000-2004)”. *South African Journal of Economics*, 74(4), 2006.
- **Baigrie, N., & Eyal, K.** An evaluation of the determinants and implications of panel attrition in the National Income Dynamics Study (2008-2010) *South African Journal of Economics* Vol. 82:1 March 2014
- **Maluccio, J.** Using quality of interview information to assess nonrandom attrition bias in developing-country panel data. *Review of Development Economics*, 8(1): 91-109, 2004

Additional data

- studying status in 2012 from the HEMIS data (not UWC)
- This match was performed by DHET and de-identified data returned to us. As a result we can only calculate the share of graduates studying in 2012 by WCGDS response, institution and field of study.
- Finally for the UCT subsample, we have additional institutional information and matched address code information to the Census 2011.

Neighborhood characteristics by response status – UCT graduates only

	Non Responder	Responder	Responder - weighted
Proportion of Census 2011 SAL:			
Employed	0.70	0.7	0.7
Household annual income:			
<614000	0.84	0.83	0.83
614000-2457600	0.15	0.15	0.15
>2457600	0.02	0.02	0.02
Small households	0.94	0.94	0.94
Large households	0.06	0.06	0.06
No schooling	0.01	0.01	0.01
Some Primary	0.11	0.11*	0.11
Completed Primary	0.02	0.02**	0.02*
Some secondary	0.17	0.17**	0.17
Grade 12	0.24	0.24	0.24
Higher	0.34	0.35*	0.34
Other	0.00	0	0
English	0.59	0.6	0.59
Afrikaans	0.20	0.2	0.2
African language	0.14	0.13	0.13
Married	0.45	0.45	0.45
African	0.23	0.22	0.23
Coloured	0.19	0.19	0.2
Indian	0.07	0.06	0.06
White	0.54	0.55	0.54
Individual annual income:			
<51200	0.80	0.8	0.8
51201-204800	0.04	0.04	0.04
>204801	0.01	0.01	0.01
Sample size	3643	1068	1068

Qualification type by Campus

Qualification type	Institution				Total
	CPUT	UCT	US	UWC	
Certificate/diploma	61.36	4.67	1.48	11.12	21.76
Postgraduate certificate	0	14.08	19.65	8.59	10.68
Bachelor's	35.79	48.26	44.09	53.14	43.99
Honours	1.42	14.03	17.25	16.86	11.62
Master's	1.28	16.37	15.18	8.73	10.32
Doctorate	0.15	2.6	2.36	1.56	1.63
Total	100	100	100	100	100

No information on why people did not respond

- No information on who was tried on the phone and the reason for not responding
- No information on how many emails bounced back etc

- Differences in characteristics does not necessarily imply bias for the analysis
- We use an selection probit to test for random response in relation to employment