Why weights do not matter, but do harm

Harry BG Ganzeboom

Department of Sociology, VU University Amsterdam

Presentation for the ISSP Research Session

Reykjavik [IS]. April 30 2023

Citation / Earlier presentations

Please cite as: Ganzeboom, Harry BG. (2023). "*Why weights do not matter, but do harm*" [Presentation]. Amsterdam: Department of Sociology VUA. http://www.harryganzeboom.nl/pdf/index.htm.

Earlier presentations:

- ISA Forum, Porto Alegre (BR), July 2022 (online)
- SILC Seminar, Department of Sociology, VU University Amsterdam, June 2 2020.
- ISSP Research Session, Reykjavik, April 29 2019 (Cancelled).

At the beginning of their textbook's section on weighted estimation of regression models, Angrist and Pischke (2009, p. 91) acknowledge, "Few things are as confusing to applied researchers as the role of sample weights. Even now, 20 years post-Ph.D., we read the section of the Stata manual on weighting with some dismay." After years of discussing weighting issues with fellow economic researchers, we know that Angrist and Pischke are in excellent company.

Solon et al., 2013, What are we weighting for? p.301

Motivation

- ISSP is making “design weights” an important issue and urges its members to collect precise information on the number of units from which the respondents are selected (= household size).
- The weight will then be proportional to how many potential cases a single respondent will “represent”.
- ***This is a bad idea*.**

GENERAL CONCLUSIONS

- Weights do not make a difference to point estimates, because:
  - Ingredients of WEIGHT are often among the predictor set of regression models: Gender, Age, Education.
  - Ingredients of WEIGHT are unrelated to the dependent variable: Region, Household Size
- Weighted estimation harms statistical precision (SE), which can be quantified as a loss of 20% - 40% of the data.
- So:
  - Do not use weights
  - Alternative: include WEIGHT in the predictor set of your model.

What are weights?

- **Post-stratification weights**: adjust sample distributions to the relative frequency of categories of some (population) standard.
- **Efficiency (or: design) weights**: correct for the (in)efficiency in sampling designs, arising from departures of SRS (simple random sampling), most typically by multi-stage clustered or stratified sampling.
- Population weights: adjust to the absolute frequencies in some population.

Post-stratification weights

- Construction: (A) IPF (iterative proportional fitting) or (B) by taking the inverse of a logistic regression that predicts sample membership (response) from a larger sampling frame.
- Predictors used most often: gender, age, region, marital status, education, employment status.
- Not often used: voting in last election, party preference in last election, although these are the only characteristics for which we know (A) the exact population values – and (B) that they affect survey results.
- Post-stratification weights typically have an average of 1.0 → they do not affect total N.
- Often recommended: avoid weights outside 0.3 / 3.0 range (which implies a 1/10 odds). ***This is good advice.***

Two kinds of post-stratification weight criteria

- It is useful to distinguish the weighting criteria into two categories:
  - Variables that have NO association with the outcomes. This is often the case for region, gender, marital status. The weighting by these variables do not affect the point estimates.
  - Variables that DO have an association with the outcomes. This is most often the case for age and education. However, weighting by these variables does not affect point estimates when these variables are part of the predictor set.
- ➔ ***Post-stratification weights should never make a difference to point estimates!!***
- Potential exception: univariate statistics such as national (unconditional) means.

The hidden injury

- Weights mostly often do not make a difference to the point estimates, but they do (and should) affect inferential statistics: SE, CI, testing.
- No wonder, ***weighting is a form of duplicating data.***
- The hidden injury arises for weights that do and for weights that do not affect point estimates.
- Such effects can be taken into account by statistical procedures that correct for clustering (e.g. **`svy`** in Stata, Complex Sampling in SPSS).
- ***They are horribly strong.***

Weighting = data duplication

- Weighting is a form of data duplication:
  - If weight > 1, duplicate existing cases
  - If weight < 1, remove existing cases.
- If conceptualized as data duplication, weighting does not sound like a good idea.
- And it should certainly not go unpunished: your **effective sample size** should be adjusted.
- This is what **svy** (Stata) and Complex Sampling (SPSS) do for you.

Efficiency (design) weights

- Efficiency weighting is much more important than post-stratification.
- Efficiency weights should NOT be 1.0 on average.
  - Typically, efficiency weights are below 1.0, if the sampling design is less efficient than SRS (due to clustering or post-stratification).
  - (Efficiency weights can be above 1.0 if the design is more efficient than SRS, which can happen with stratified sampling.)
- However, efficiency weights can only be calculated in a certain context with specified variables. The effect of clustering / stratification / post-stratification depends upon the dispersion of these variables over your sampling units.
- For efficiency weighting it is necessary that information about the clustering / strata units is present in the dataset. Household size is just one of these.
- ***There is a lot of nonsense about 'design weights' on the ESS website.***
  - ***ESS design weights are actually post-stratification weights,***
  - ***and they should not be used.***

Design effects

- In sampling theory (Kish, 1965) the effect of sampling designs is commonly expressed in the design effect, defined as:

  deff = (variance of estimate in SRS / variance of estimate in actual sampling design).

- Interpretation of ***deff***: how many more cases do you need to obtain the same precision (confidence interval) as in SRS?
- Effects of post-stratification weights can also be expressed in ***deff.***

TEST ON ESS R1-R9

Weights in ESS

- ESS calls post-stratification weights “design weights”: **`pspwght`**.
- They have mean 1.00 (or abouts) within each country and ESS-round.
- Eight countries in ESS never provided weights or weights with very minor variations. (I have excluded these countries from the analysis.)
- Some countries have provide weights in some rounds but not in others.
- Average SD(**`pspwght`**) = 0.490. Min: 0.016 Max: 6.207.

Testcase: Religiosity

- I take Religiosity in ESS as a testcase. It is an index of three indicators: **mean(-zrlgdgr, zrlgatnd, zpray)**. Scaled as zRelig (M=0, SD=1).
- Religiosity:
  - Strongly differentiated by country
  - Strongly predicted by age, education, gender – criteria that have likely been used to construct the **pspwght**.

Test #1: individual level models

- How do individual level predictors of religiosity behave, with and without weighting?
- Expectations:
  - No change of B's
  - Larger SE's, if weighting is taking into account with **svy** estimation (in Stata).

Determinants of **pspwght**

Pearson correlations

	pspwgh	age	educyr	female	Zrelig
pspwgh	1.000	-.116	.001	-.018	.020
age	-.116	1.000	-.153	.020	.153
educyr	.001	-.153	1.000	-.004	-.124
female	-.018	.020	-.004	1.000	.165
zRelig	.020	.153	-.124	.165	1.000

Standardized regression

	Beta	t
(Constant)		275.6
age	-.124	-65.5
educyr	-.013	-6.9
female	-.022	-11.8
zRelig	.041	21.6

adj R2=1.6%

Table 2: Unweighted and weighted estimates of determinants of religious involvement

	Unweighted				weighted				
	Model 1a		Model 1b		Model 2				
X	B	SE	B	SE	B	SE	SE(svy)	SE(svy) / SE	deff
Essround	-2.013	0.072	-2.023	0.072	-2.144	0.072	0.080	1.113	1.238
Age	1.048	0.011	1.082	0.011	1.022	0.011	0.012	1.131	1.279
Educyr	-1.371	0.046	-1.348	0.046	-1.349	0.046	0.054	1.166	1.359
Female	31.43	0.330	31.57	0.330	30.481	0.329	0.370	1.122	1.259
Weight			8.781	0.337					
Adj R2	23.17%		23.35%		22.83%				

ESS R1-9, 30 countries with poststratification weights. Age 18-74. N=288951. Y is zscore multiplied by 100.

All models include country - fixed effects

Results for ESS

- Most effects and adj.R2 are smaller for weighted estimates – this would unlikely be the case if weighting would improve the model.
- SE's increase for weighted **svy** estimates – in different ways for different variables.
- ***deffs*** (average: 1.28) suggest that we effectively lose about 28% of the cases by post-stratification. For education this is 36%!

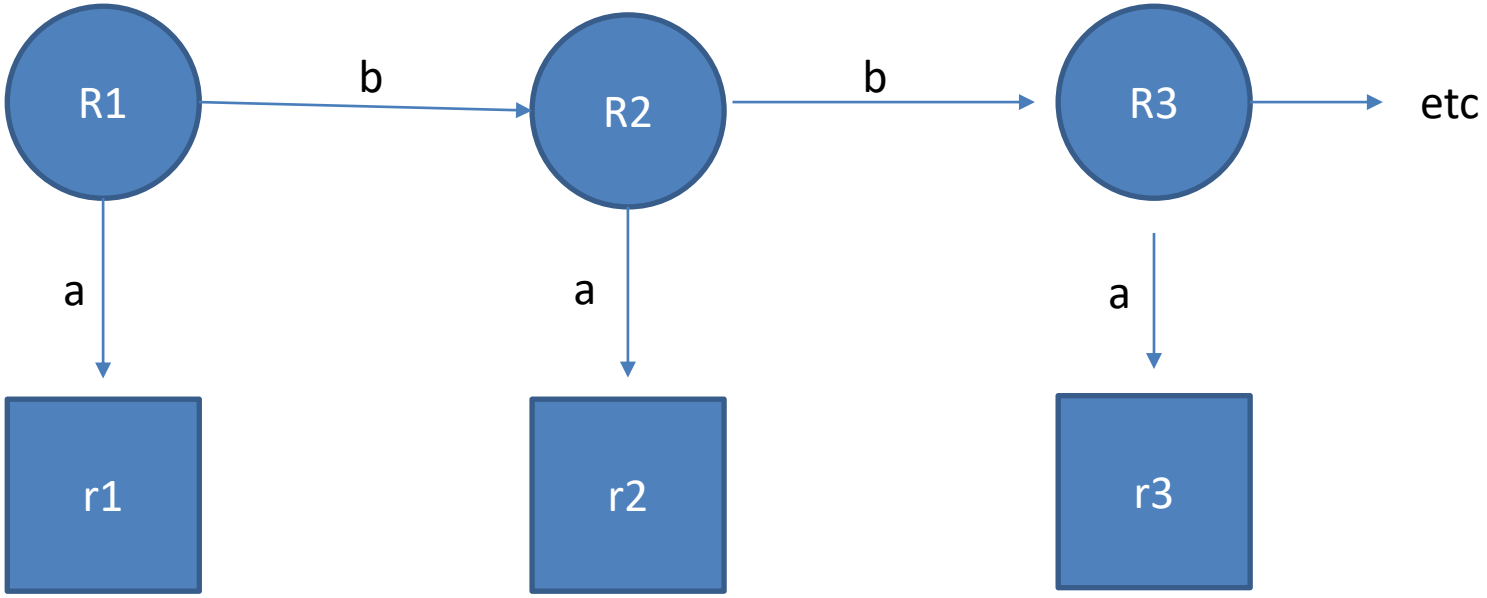
➔ ***Weighting makes no difference, but does harm!***

TEST #2: aggregate data

Another way to test: simplex model on unconditional means

- The most plausible effect of weighting is on unconditional means.
- However, weights might make the unconditional means more valid ('representative'), but also less reliable.
- A way to test this is the simplex model on country means. This can separate true change from random instability.

Simplex model



b: true stability; a: measurement reliability; model can be estimated on three waves, but better if there are nine waves.

Simplex (markow): R1 does not affect R3 (etc) unless via R2.

Expectations and complications

- **b** (true stability) should be the same for the unweighted and weighted country means.
- **a** should be larger for weighted means if these are “more representative”, but smaller for the weighted means if weighting brings in random variations.
- Simplex models are hard to estimate: (A) few and incomplete cases (N-available = 30); (B) Correlations are extremely strong, but indicate some true change as they taper off from the diagonal.

Correlations unweighted means									
	zrelig1	zrelig2	zrelig3	zrelig4	zrelig5	zrelig6	zrelig7	zrelig8	zrelig9
zrelig1	1	.993	.990	.965	.971	.960	.959	.947	.952
	16	15	11	14	14	13	13	14	12
zrelig2	.993	1	.978	.969	.954	.952	.962	.952	.941
	15	19	13	16	15	15	12	14	12
zrelig3	.990	.978	1	.981	.971	.958	.975	.966	.944
	11	13	18	18	16	15	11	12	12
zrelig4	.965	.969	.981	1	.986	.968	.974	.964	.948
	14	16	18	24	21	18	14	15	13
zrelig5	.971	.954	.971	.986	1	.984	.980	.977	.976
	14	15	16	21	21	18	14	15	13
zrelig6	.960	.952	.958	.968	.984	1	.991	.988	.986
	13	15	15	18	18	22	13	16	13
zrelig7	.959	.962	.975	.974	.980	.991	1	.995	.992
	13	12	11	14	14	13	14	14	11
zrelig8	.947	.952	.966	.964	.977	.988	.995	1	.996
	14	14	12	15	15	16	14	17	12
zrelig9	.952	.941	.944	.948	.976	.986	.992	.996	1
	12	12	12	13	13	13	11	12	15

Correlations weighted means									
	zrelig1w	zrelig2w	zrelig3w	zrelig4w	zrelig5w	zrelig6w	zrelig7w	zrelig8w	zrelig9w
zrelig1w	1	.991	.993	.975	.975	.962	.964	.950	.962
	16	15	11	14	14	13	13	14	12
zrelig2w	.991	1	.981	.971	.961	.966	.967	.948	.944
	15	19	13	16	15	15	12	14	12
zrelig3w	.993	.981	1	.979	.967	.961	.979	.961	.944
	11	13	18	18	16	15	11	12	12
zrelig4w	.975	.971	.979	1	.984	.964	.976	.960	.946
	14	16	18	24	21	18	14	15	13
zrelig5w	.975	.961	.967	.984	1	.981	.986	.979	.977
	14	15	16	21	21	18	14	15	13
zrelig6w	.962	.966	.961	.964	.981	1	.990	.983	.987
	13	15	15	18	18	22	13	16	13
zrelig7w	.964	.967	.979	.976	.986	.990	1	.994	.991
	13	12	11	14	14	13	14	14	11
zrelig8w	.950	.948	.961	.960	.979	.983	.994	1	.997
	14	14	12	15	15	16	14	17	12
zrelig9w	.962	.944	.944	.946	.977	.987	.991	.997	1
	12	12	12	13	13	13	11	12	15

Results for the simplex model

- Estimates of true stability coefficient (**b**):
  - Unweighted 0.995 t = 3.42 (H0=1.0)
  - Weighted 0.996 t = 3.12 (H0=1.0)

  → ***Virtually identical***
- Estimates of measurement coefficient (a):
  - Unweighted **0.996** t = 2.77 (H0=1.0)
  - Weighted 0.992 t = 3.42 (H0=1.0)

  → ***Weighting creates more noise than it repairs***

Substantive conclusions on ESS R1-R9

- Difference between weighted and unweighted means is almost nothing.
- If statistically evaluated:
  - Difference is statistically significant (N = 30 countries!)
  - The weighted estimates have (slightly) larger error variance than the unweighted estimates.

→ Weights do not make a difference, but do harm!

TEST ON ISSP 2009-2018

Test on ISSP 2009-2018: ATTEND

- ATTEND is a good indicator to test the effects of weighting:
  - It is available in all ISSP waves
  - It varies by:
    - Countries
    - Year of survey (linear trend)
    - Individual determinants: gender, age, education.
    - Other ingredients of WEIGHT: Region.
- If weights have effects on results, ATTEND is a good indicator to analyse them.

Test #1: Individual level models

Models

A. OLS regression, unweighted
B. OLS regression, weighted
C. OLS regression, weighted with Stata **svy.**

- Comparison of B and C defines *deff: how much power (cases) do you loose by weighted estimation: SE(A) / SE(B)?*
- Complete cases: N=18. Available cases: N=50.

Determinants of **WEIGHT**

Correlations

	WEIGHT	ATTEND	YEAR	FEMALE	AGE	DEGREE
WEIGHT	1.000	.023	-.002	-.072	-.149	-.005
ATTEND	.023	1.000	.045	-.088	-.040	.123
YEAR	-.002	.045	1.000	-.010	.040	.099
FEMALE	-.072	-.088	-.010	1.000	.006	-.022
AGE	-.149	-.040	.040	.006	1.000	-.173
DEGREE	-.005	.123	.099	-.022	-.173	1.000

Standardized Regression

	Beta	
ATTEND	.015	10.4
YEAR	.007	4.6
FEMALE	-.070	-48.2
AGE	-.154	-104.9
DEGREE	-.036	-24.1

Table 4: Unweighted and weighted estimates of determinants of zATTEND

	Unweighted		Weight as control		Weighted		SVY estimation			
	B	SE	B	SE	B	SE	B	SE	SE(3)/ SE(1)	DEFF
FEMALE	-0.1522	0.00254	-0.1513	0.00254	-0.1557	0.00252	-0.1557	0.0029	1.14	1.30
ZDEGREE	-0.0058	0.00142	-0.0056	0.00142	-0.0072	0.00142	-0.0072	0.0018	1.23	1.52
ZAGE	-0.1179	0.00133	-0.1170	0.00135	-0.1184	0.00133	-0.1184	0.0016	1.18	1.39
YEAR	0.1819	0.00414	0.1818	0.00414	0.1921	0.00414	0.1921	0.0048	1.16	1.34
WEIGHT			0.0110	0.00230						
adj R2	26.36%		26,36%		25.93%		25.94%			

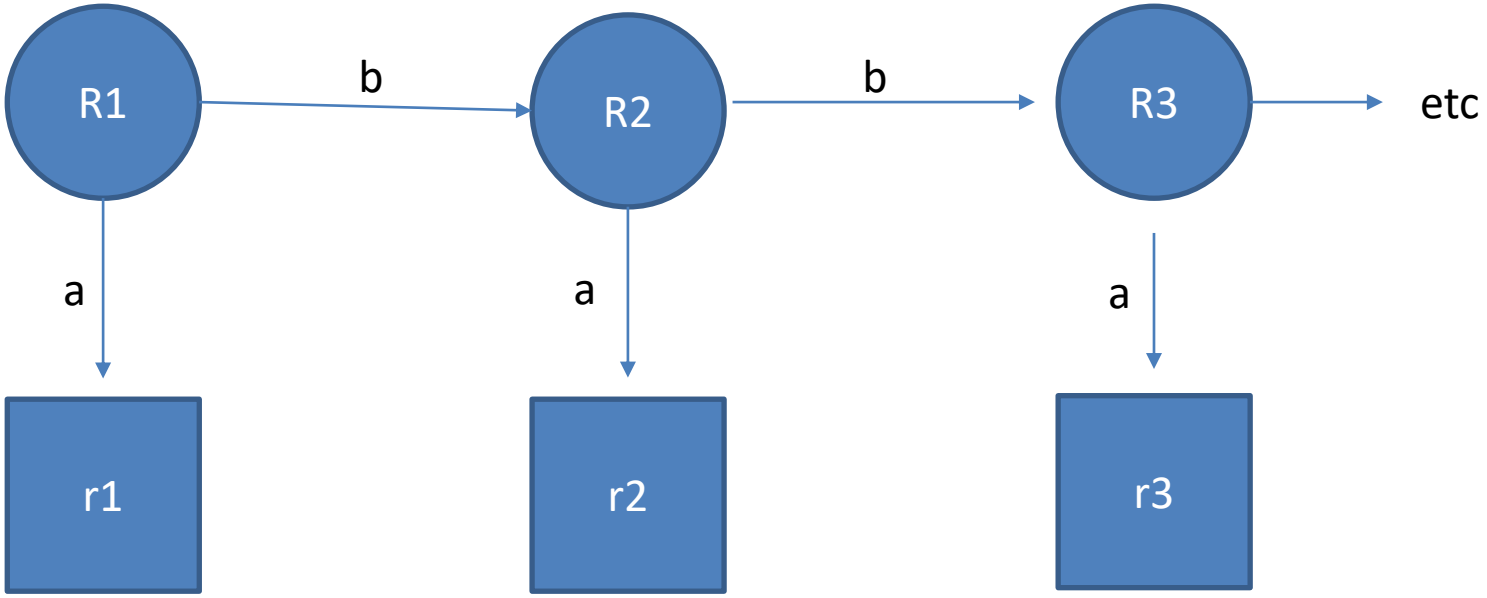
Results ISSP

- Most effects and adj.R2 are smaller for weighted estimates – this would unlikely be the case if weighting would improve the model.
- SE's increase for weighted **svy** estimates – in different ways for different variables.
- ***DEFFs*** (average: 1.45) suggest that we effectively loose about 45% of the cases by post-stratification. For education this is 52%!

Test #2: Simplex model on unconditional means

- The most plausible effect of weighting is on unconditional means.
- However, weights might make the unconditional means more valid ('representative'), but also less reliable.
- A way to test this is the simplex model on unconditional (country) means. This can separate true change from random instability.

Simplex model



b: true stability; a: measurement reliability; model can be estimated on three waves, but better if there are nine waves.

Simplex (markow): R1 does not affect R3 (etc) unless via R2.

Table 2: Correlations between mean ATTEND scores of 50 ISSP countries 2009-2018

Correlation Matrix Unweigted Means										
	att2009	att2010	att2011	att2012	att2013	att2014	att2015	att2016	att2017	att2018
att2009	1.000	.963	.979	.973	.940	.950	.950	.944	.973	.958
att2010	.963	1.000	.965	.964	.967	.915	.919	.960	.960	.977
att2011	.979	.965	1.000	.993	.971	.985	.988	.965	.979	.954
att2012	.973	.964	.993	1.000	.976	.974	.981	.952	.971	.950
att2013	.940	.967	.971	.976	1.000	.967	.951	.976	.956	.951
att2014	.950	.915	.985	.974	.967	1.000	.984	.962	.938	.910
att2015	.950	.919	.988	.981	.951	.984	1.000	.953	.958	.911
att2016	.944	.960	.965	.952	.976	.962	.953	1.000	.978	.951
att2017	.973	.960	.979	.971	.956	.938	.958	.978	1.000	.962
att2018	.958	.977	.954	.950	.951	.910	.911	.951	.962	1.000

Correlation Matrix Weighted Means

	attw2009	attw2010	attw2011	attw2012	attw2013	attw2014	attw2015	attw2016	attw2017
2009	1.000	.961	.983	.976	.945	.947	.948	.946	.974
2010	.961	1.000	.964	.960	.968	.909	.916	.960	.961
2011	.983	.964	1.000	.994	.971	.983	.986	.966	.980
2012	.976	.960	.994	1.000	.976	.969	.978	.953	.975
2013	.945	.968	.971	.976	1.000	.968	.954	.982	.958
2014	.947	.909	.983	.969	.968	1.000	.987	.965	.940
2015	.948	.916	.986	.978	.954	.987	1.000	.961	.959
2016	.946	.960	.966	.953	.982	.965	.961	1.000	.978
2017	.974	.961	.980	.975	.958	.940	.959	.978	1.000

N of Countries										
att2009	41	29	27	33	28	30	33	32	27	31
att2010	29	32	24	30	25	26	27	27	23	27
att2011	27	24	29	27	24	27	24	26	21	25
att2012	33	30	27	39	30	32	32	31	27	29
att2013	28	25	24	30	34	29	31	30	26	27
att2014	30	26	27	32	29	35	32	33	27	30
att2015	33	27	24	32	31	32	38	34	31	30
att2016	32	27	26	31	30	33	34	37	29	32
att2017	27	23	21	27	26	27	31	29	32	28
Att208	31	27	25	29	27	30	30	32	28	35

Results for the simplex model

- Estimates of true change coefficient (**b**):
  - Unweighted 0.996
  - Weighted 0.996

  → Virtually identical
- Estimates of measurement coefficient (**a**):
  - Unweighted **0.986**
  - Weighted 0.987
- Fit:
  - L2 Unweighted 175.9 / 53
  - L2 Weighted 194.3 / 53

  ➔ ***Weighting creates as much noise than it repairs!!***

Conclusions on aggregate model

- Difference between weighted and unweighted means is almost nothing.
- If statistically evaluated:
  - Difference is NOT statistically significant (N = 50 countries!)
  - ~~The weighted estimates have (slightly) larger error variance than the unweighted estimates.~~

→ Weights do not make a difference, but do harm!

GENERAL CONCLUSIONS

- Weights do not make a difference to point estimates, because:
  - Ingredients of WEIGHT are often among the predictor set of regression models: Gender, Age, Education.
  - Ingredients of WEIGHT are unrelated to the dependent variable: Region.
- Weighted estimation harms statistical precision (SE), which can be quantified as a loss of 20% - 40% of the data.
- So:
  - Do not use weights
  - Alternative: include WEIGHT in the predictor set of your model.

WHAT TO DO?

What if weighting makes a difference?

- This rarely happens…
- But if is happens, you are in real trouble. What do the results mean? Are they any better than the unweighted results?

How to handle post-stratification weights

- Strategy 1: always analyze unweighted data. Check whether weighting would make a difference to your point estimates. If not, report unweighted results. If yes, resort to strategy 2 or 3.
- Strategy 2: Use the weight as a control variable in your model.
- Strategy 3: Calculate unstandardized regressions. Calculate weighted means and SD's for all variables. Calculate standardized coefficients using weighted SD.
- ***Send me $ 1 (using Skype money transfer to h.ganzeboom) every time Strategy 1 works.***

References

- Kish, Lesley. 1965. *Survey Sampling*. New York, NY: Wiley.
- Solon, Gary, Steven J. Haider, and Jeffrey M. Wooldridge. 2015. “What Are We Weighting For?” *Journal of Human Resources* 50 (2): 301–16. https://doi.org/10.3368/jhr.50.2.301.
- Wikipedia Team (2023) https://en.wikipedia.org/wiki/Design_effect