Placebo Statements in List Experiments

Kai Ostwald* Guillem Riambau†

May 9, 2017

Abstract

List experiments are a widely used survey technique for estimating the prevalence of socially sensitive attitudes or behaviors. This paper shows a potential vulnerability in their design: in a conventional list experiment, the treatment group respondents see a greater number of items ($N + 1$) than the control group respondents (N). This can lead to mechanical inflation in the mean response of the treatment group, simply because respondents see a greater number of items. We test this through a *placebo* control statement that is by design false for all respondents. Using original data, we find the $N + 1$ item placebo group mean to be greater than N standard control group mean, even though there should be no difference in means between the groups. This clearly indicates the presence of mechanical inflation. Further, we show that this effect is most pronounced among elderly and low socioeconomic status respondents. We recommend using the essentially costless placebo statement in list experiment control groups as a preventive measure against potentially upward biased findings.

Keywords: List experiments, item count technique, survey design, placebo statement, respondent fatigue.

JEL classification: C81, C83

SSRN: `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2965560`

*University of British Columbia. email: kai.ostwald@ubc.ca

†Department of Social Sciences, Yale NUS College. email: guillem.riambau@yale-nus.edu.sg

Research carried out under IRB A-16-032E from the National University of Singapore and IRB H16-01592 from the University of British Columbia. We would like to thank Eugene Choo, Allen Hicken, Gillian Koh, Edmund Malesky, Steven Oliver, and the audiences at NUS, Yale-NUS, Universitat Autònoma de Barcelona, and the WPSA and AAS conferences, for insightful comments and suggestions. All errors are ours.

1 Introduction

List experiments (also known as item count technique) are a widely-used survey technique designed to elicit true preferences on sensitive topics (Rosendfeld et al. (2016)). They work as follows: a sample of respondents is divided into control and treatment groups. The control group is shown N non-sensitive statements and asked to indicate how many are true. The treatment group is shown $N + 1$ statements, where the N statements are the same as the control group, but the $+1$ is a sensitive item that may elicit socially desirable responses if asked directly. The difference in the mean number of true statements between the control and treatment groups is interpreted as the percentage of the sample for whom the sensitive statement is true. This technique has been used to estimate the prevalence of a wide-range of socially sensitive attitudes and behaviors, from prejudice against ethnic minorities, to sexual practices and voting behavior.[1]

This paper discusses a fundamental but largely overlooked weakness of conventional list experiments and offers a simple preventative fix.[2] In a conventional list experiment, the treatment group receives $N + 1$ statements; in other words, one statement more than the control group. This introduces the risk of a mechanical bias in which treatment group respondents indicate a higher number of true statements *simply because they see a greater number of total statements* than the control group. If present, this mechanical bias would increase the mean number of true statements in the treatment group, thereby artificially inflating the treatment effect and biasing towards significant findings.

We test for this through an original survey ($N = 1,151$)[3] with an embedded list experiment that includes two control groups: the standard control group has 4 statements, while the *placebo control group* receives $4 + 1$ statements, where the $+1$ is a *placebo statement* that is by design false for all respondents. Our findings reveal the presence of

[1] See Koisuke Imai's webpage for a list of examples: `http://imai.princeton.edu/research/files/listExamples.pdf`.

[2] To the best of our knowledge, only Kramon and Weghorst (2012) and Ahlquist et al. (2014) have looked into this: we discuss their findings in section 4 (concluding remarks). For papers that discuss the advantages and disadvantages of list experiments, see Rosendfeld et al. (2016), Glynn (2013), Blair et al. (2014), or Tsuchiya et al. (2007).

[3] The broader survey examined community attitudes and voting behavior. See section 2 or Ostwald and Riambau (2017) for full details.

a mechanical inflation in the placebo group. The effect is strongly heterogeneous: it is pronounced in lower socio-economic status (SES) and elderly respondents, but absent in their counterparts.

Our survey makes relatively low cognitive demands on respondents: it is comprised only of closed answers and requires less than 10 minutes to complete. We expect that the mechanical inflation we observe would be a greater threat in longer and more cognitively demanding surveys, where respondents are more likely to resort to *satisficing*, i.e., skipping one or more cognitive steps in order to fully answer a question, making only the minimal effort required to produce a plausible response (Krosnick (1999); Kramon and Weghorst (2012)). In short, the findings presented here suggest that conventional list experiments may over-report agreement with the sensitive item due to a simple behavioral bias. Given this, we recommend that future list experiments include a placebo statement in the control group: this is an essentially costless protective measure, as it does not increase cognitive demands or alter interpretation of survey experiments, but does protect against the observed mechanical bias.

2 Data and survey design

The data described below come from a list experiment embedded in a survey on social and political beliefs in Singapore, conducted between September 2016 and April 2017.[4] The survey was administered in person by a multi-ethnic team of enumerators comprised of local university students. It was available in English, Chinese, and Malay. We used a probability proportional to size strategy to produce a representative sample.[5] Singapore has among the world's highest levels of educational attainment and close to full literacy.[6] The survey took most respondents between 5 and 10 minutes to complete. While we

[4]For full details on the survey methodology and the scope of the other projects, we refer the reader to Ostwald and Riambau (2017). A copy of the survey we used is included in the supplementary materials.

[5]As in Wong (2013), we include only public housing residents in our study, primarily due to private housing being inaccessible to enumerators. In Singapore, over 80% of the population live in public housing, so this strategy does not produce any significant bias. A balance table is included in the supplementary materials.

[6]CIA World Factbook, available at `https://www.cia.gov/library/publications/the-world-factbook/geos/sn.html`, last accessed on April 19, 2017.

collected 3,335 total responses, only the 1,151 responses in the control groups are relevant to this paper.

The list experiment was designed to measure beliefs on the secrecy of the ballot in Singapore. We opted for two control groups, into which respondents were randomly assigned. The *standard control group* received 4 neutral statements, while the *placebo control group* received 4 neutral statements plus 1 (necessarily false) placebo statement. The treatment groups, which we do not address further in this paper, likewise received $4+1$ statements, where the $+1$ statement was a sensitive item. A balance table, included with the supplementary materials, does not show any signs of bias between the groups on observables.

All groups received the same instructions: "Look at the following statements below. Can you tell us how many statements are true for you? **Please don't tick individual statements, just tell us the total number"** [Emphasis in the original].

The four neutral statements are as follows:

- I like durian very much.
- I have had a haircut within the last four days.
- I have eaten at a hawker centre at least once in the last two days.
- I have gone to a local "Meet the People" session at least once in the last year.

These statements were chosen following the generally accepted criteria for list experiments: they fit naturally into the context of the survey, they are uncorrelated (both with one another and with other broader socio-economic characteristics), and they are resistant to ceiling effects.

The placebo statement is designed to be false for all respondents. It is as follows:

- I have been invited to have dinner with PM Lee at Sri Temasek next week.[7]

The placebo statement is the equivalent of being asked to have dinner with the President of the United States in the White House or some other equally improbable event. We

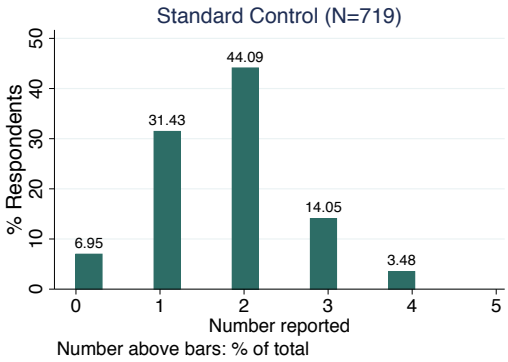
[7]Primer Minister of Singapore, Lee Hsien Loong, since 2004. Sri Temasek is the official residence of the Prime Minister.

assume that it is obviously false for all respondents and that it should easily be recognized as false. In total, the standard control group has 719 observations, relative to 432 for the placebo group.

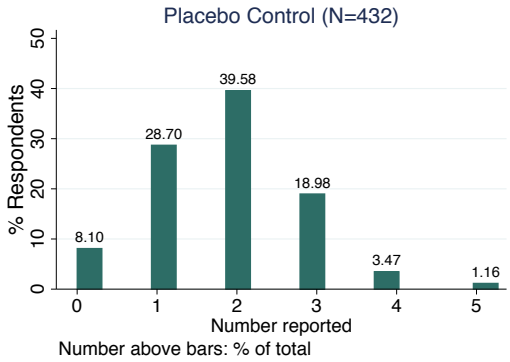
3 Results

Figure 1 shows the frequency distribution of reported true statements across the standard and placebo control groups. Notable is that few respondents indicate 0 or 5 statements in the placebo group, which suggests that the presence of a clearly false placebo statement does not induce respondents to indicate extreme counts. Rather, the clearest inflation in the placebo group is with 3 reported true statements. This may be a reflection of the perceived midpoint shifting to 3.

Figure 1: Distribution of reported true statements for each group



(a) Frequencies: Standard control group



(b) Frequencies: Placebo control group

Table 1 provides a more detailed summary of the overall findings. For the whole sample, the mean number of reported true statements is higher (1.84) for the placebo group than for the standard control group (1.76). While the .08 difference in means falls just short of the 5% level of significance, its magnitude is substantial: a simplified interpretation suggests that the inclusion of the +1 placebo statement induces roughly 8% of respondents to increase their reported true statements by 1 above their counterparts in the standard control group.

Table 1: Mean number of reported true statements, by group. In parentheses, number of observations.

		Control		p-value
		Standard	Placebo	
Whole sample		1.76 (719)	1.84 (432)	0.0624
Political knowledge	Low	1.77 (458)	1.90 (263)	0.0393
	High	1.75 (259)	1.76 (169)	0.4449
Education	None or primary	1.48 (97)	2.07 (46)	0.0007
	Secondary	1.86 (186)	1.95 (111)	0.2095
	Post-secondary	1.78 (423)	1.77 (262)	0.5459
Household income	$<$ \$3.5K per month	1.76 (261)	1.96 (162)	0.0251
	$\geq$ \$3.5K per month	1.81 (378)	1.79 (228)	0.5914
Age	61+	1.56 (144)	1.86 (82)	0.0124
	$18-60$	1.80 (562)	1.84 (346)	0.2680

$p-$values for one-sided test of difference in means between the control and the placebo. E.g., 0.0393 is the p-value resulting of testing the difference in means between low political knowledge control group and low political knowledge placebo group. Political knowledge: '1' if respondent knows the electoral district in which they reside, '0' if answer is left blank, wrong answer is given, or respondent ticked "I know, but I cannot remember" or "I don't know". Sample: all respondents (i.e., Singaporean citizens, permanent residents, and temporary residents).

Table 1 also reports mean number of true statements by subgroups on the dimensions of political knowledge, educational attainment, household income, and age. We opt for simple categories to facilitate comparisons: respondents are coded as having high political knowledge when they are able to correctly name their electoral district; household income is above and below 3,500 Singapore dollars per month (which represents roughly the bottom third), while age is above or below 60 years.

The findings suggest that the treatment effect of the placebo statement is highly heterogeneous: for the politically knowledgeable, relatively educated, middle and upper income, and younger subgroups, the difference in means between the standard and placebo control groups is insignificant, meaning that the inclusion of the placebo statement does not inflate the reported number of true statements. By contrast, the difference in means is pronounced both in terms of magnitude and significance among the counterpart subgroups. This provides a strong initial indication of which respondent types are most vulnerable to mechanically inflating their true statement count in conventional list experiments.

We further examine the correlation between individual characteristics and propensity to inflate the placebo score using the model specifications from Imai (2011) and Blair and

Imai (2012). Specifically, we estimate the following linear regression:

$$LIST_i = \alpha + \beta X_i + \delta PLACEBO_i + \gamma\,(PLACEBO_i \times X_i) + \varepsilon_i, \tag{1}$$

where X_i are sociodemographic variables and $PLACEBO_i$ is a dummy that takes value 1 if the respondent was part of the placebo group, 0 if part of the standard control group. This is the same as expression (6) in Imai (2011).[8] γ is the vector of our coefficients of interest: we expect it to be significant for the variables specified in table 1.

Table 2 reports the results. Specifications $(1) - (4)$ confirm the unconditional results of Table 1 using fixed effects and clustered standard errors. Specifications $(5) - (9)$ further add sociodemographic controls (gender, ethnicity, and apartment size): earlier findings are not fundamentally altered. Specification (9), which includes all controls and variables of interest, reveals that educational attainment is the strongest predictor of inflating the number of true statements due to the inclusion of the (necessary false) placebo statement. Other variables (especially, income and political knowledge) likely lose their significance due to power and multicollinearity issues: the correlation between education and political knowledge is relatively large (0.21), as is the correlation between education and household income (0.49). Finally, note that the R^2s are generally quite low: this is evidence that, as expected, agreement to the statements in our list experiment is randomly distributed across the population and hard to correlate with observables.

From specification (9), we can see that, on average, a respondent with primary school education (or below) is likely to report a number 0.2 points higher than a respondent with a secondary school diploma, and 0.4 points higher than a respondent with college education. Figure 2 shows this graphically for the two most important explanatory variables: educational attainment and age. The left panels show the predicted responses and a smooth polynomial fit from respondents in the placebo group (using model predictions from specification (9) in Table 2). Right panels show the predicted responses and fit for respondents in the standard control group. We can see that, for instance, while elderly re-

[8] Expression (6) in Blair and Imai (2012).

Table 2: Estimated coefficients from the item count technique regression models where the extra statement in the treatment group is designed to be false for all respondents (placebo). Linear regression with interactions. See expression (1) for details on the specification.

	No controls				With controls				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Placebo item									
PLACEBO × ...									
... 61+ years old	0.258				0.240				0.102
	(0.124)				(0.141)				(0.170)
... Education		-0.045				-0.045			-0.039
		(0.011)				(0.011)			(0.016)
... Political knowledge			-0.191				-0.207		-0.142
			(0.130)				(0.162)		(0.162)
... Hhd. Income				-0.021				-0.023	-0.002
				(0.011)				(0.013)	(0.015)
Panel B: Control item									
61+ years old	-0.269				-0.281				-0.213
	(0.083)				(0.088)				(0.106)
Education		0.018				0.022			0.156
		(0.009)				(0.009)			(0.011)
Political knowledge			0.021				0.036		-0.024
			(0.062)				(0.060)		(0.069)
Hhd. Income				0.003				0.008	-0.004
				(0.007)				(0.008)	(0.009)
Controls	No	No	No	No	Yes	Yes	Yes	Yes	Yes
Clustered s.e.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
District fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.03	0.03	0.02	0.03	0.05	0.05	0.04	0.05	0.06
Observations	1149	1149	1149	1149	1149	1149	1149	1149	1149

Standard errors (in parenthesis), clustered at the electoral district level (24 districts in the dataset). Dependent variable: number stated in the list experiment. Education: years of schooling. 61+ years old: dummy for being 61 years of age or older. Hhd. Income: monthly household income, in thousands of Singapore dollars. Political knowledge: '1' when respondent correctly names the electoral district in which they reside; otherwise '0'. Controls: gender, ethnicity, apartment size. *PLACEBO:* Dummy for being in the placebo group. All coefficients reported in panel A are the interaction of *PLACEBO* × 'variable'.

spondents are more likely to report smaller numbers (0.3 smaller on average) than younger ones in the standard control group, this difference disappears when respondents are in the placebo control. This also suggests mechanical inflation in the elderly subgroup.

Finally, we use the non-linear specification suggested by Imai (2011) and Blair and Imai (2012) to further check for the robustness of our results.[9] We solve the model using

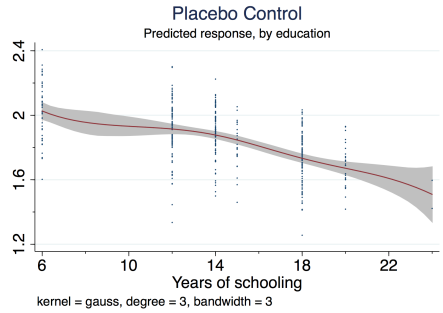
[9]Nonetheless, Blair and Imai (2012) show that when the number of observations is above 1,000, the difference in efficiency between the linear and the non-lineal models is minimal (page 63).

the **R** package `'list'`. We estimate:[10]

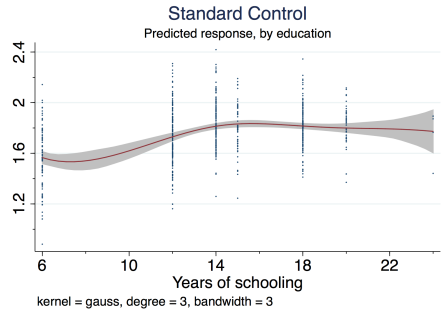
$$LIST_i = f(X_i, \beta) + PLACEBO_i \times g(X_i, \gamma) + \varepsilon_i, \tag{2}$$

where f and g are non-linear functions that represent the regression models for the conditional expectations of the control and placebo items given the covariates.

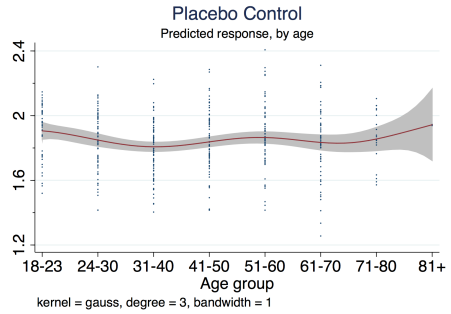
Figure 2: Predicted number of responses. Points are in-sample predictions computed using the results from specification (9) in Table 2. Line is a local polynomial fit with 95% confidence intervals.



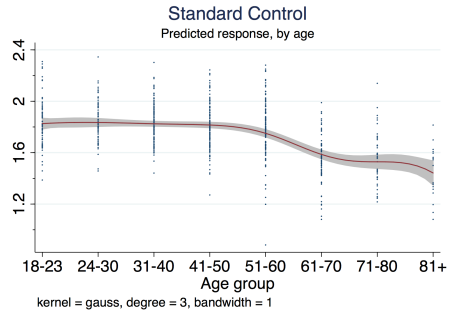
(a) Mean prediction by education (years), placebo control



(b) Mean prediction by education (years), standard control



(c) Mean prediction by age, placebo control



(d) Mean control prediction by age, standard control

The results from Table 3 provide further support for the notion that educational attainment and income predict vulnerability to mechanical inflation: both variables are significant, particularly when using the most flexible specification (unconstrained model).[11]

[10] Expression (5) in Imai (2011), expression (4) in Blair and Imai (2012).

[11] Due to convergence issues, we drop the variable 'political knowledge' and all dummies for missing observations from this specification. This accounts for the drop in observations.

Table 3: Estimated coefficients from the item count technique regression models where the extra statement in the treatment group is designed to be false for all respondents (placebo). Maximum likelihood estimation of logistic models using the `'list'` package in **R** (Blair and Imai (2012)). See expression (2) for details on the specification.

	Constrained model				Unconstrained model			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Placebo item								
Education	-0.227	-0.229	-0.144	-0.129	-0.594	-0.312	-0.267	-0.162
	(0.092)	(0.101)	(0.115)	(0.125)	(0.103)	(0.152)	(0.148)	(0.101)
61+ years old			0.954	1.019			0.332	-0.121
			(0.913)	(1.023)			(1.092)	(0.874)
Hhd. Income			-0.290	-0.292			-0.370	-0.370
			(0.196)	(0.209)			(0.345)	(0.191)
Panel B: Control item					$h_0(y; x, \psi_0)$			
Education	0.012	0.013	0.004	0.003	0.000	0.004	0.001	-0.003
	(0.008)	(0.009)	(0.011)	(0.011)	(0.010)	(0.012)	(0.011)	(0.010)
61+ years old			-0.267	-0.269			-0.162	-0.168
			(0.111)	(0.112)			(0.112)	(0.100)
Hhd. Income			-0.003	0.001			-0.003	-0.000
			(0.009)	(0.009)			(0.009)	(0.009)
					$h_1(y; x, \psi_1)$			
Education					4.154	0.082	0.224	0.501
					(591.01)	(0.080)	(0.222)	(0.274)
61+ years old							-0.935	-1.105
							(1.010)	(1.235)
Hhd. Income							0.489	1.575
							(0.355)	(0.780)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Log-likelihood	-1522.9	-1498.4	-1356.7	-1336.1	-1517.8	-1496.1	-1352.6	-1328.3
Observations	1123	1106	1007	994	1123	1106	1007	994

Standard errors in parenthesis. Dependent variable: number stated in the list experiment. Education: years of schooling. 61+ years old: dummy for being 61 years of age or older. Hhd. Income: monthly household income, in thousands of Singapore dollars. Controls: gender, ethnicity, apartment size. 'Political knowledge' not included due to convergence issues. Constrained model: control group parameters are constrained to be equal. See `http://imai.princeton.edu/software/list.html` and `http://imai.princeton.edu/talk/files/UCI11.pdf` for further details on the R package `'list'`. Webpages last accessed April 19, 2017.

4 Concluding Remarks

This paper demonstrates an inherent vulnerability of conventional list experiments: the greater number of statements in treatment groups can mechanically and artificially inflate the number of reported true statements, in particular among elderly and low-SES

respondents. This is in line with two existing studies: Kramon and Weghorst (2012) show that respondents unaccustomed to cognitively demanding tasks have a propensity to report inaccurate responses in list experiments. Furthermore, Ahlquist et al. (2014) include a placebo statement ("I have been abducted by an alien") in an online list experiment conducted in the United States; the mean of the placebo control group is approximately .06 higher than the standard control group, also suggesting the presence of mechanical inflation.

Our findings build on these two pieces by demonstrating that the prevalence of mechanically inflating the reported number of 'true' statements is heterogeneous: there is no evidence for it among high-SES and younger respondents, but it is strongly pronounced among low-SES and elderly respondents. In the latter groups, the magnitude is easily sufficient to produce incorrect conclusions about the significance of treatment effects in sensitive analyses. We do not believe that our findings are an artifact of our sample or instrument. To the contrary, Singapore is fully urbanized and industrialized, and has among the highest educational attainment rates in the world. In addition, the survey instrument could be completed in 5 to 10 minutes for most respondents and was comprised of simple, closed answers. Given the findings, longer surveys administered to respondents with lower levels of formal education could well increase mechanical inflation, particularly if the list experiment appears after respondent fatigue sets in.

We suggest a simple preventative measure: including a placebo statement (that is by design false for all respondents) in the control group does not reduce the efficacy or interpretation of list experiments, but does effectively remove the risk of mechanical inflation bias by equalizing the number of statements in control and treatment groups.

References

Ahlquist, John S., Kenneth R. Mayer, and Simon Jackman: *Alien Abduction and Voter Impersonation in the 2012 U.S. General Election: Evidence from a Survey List Experiment*, Election Law Journal, **13 (4)**, (2014), 460-475.

Blair, Graeme, and Kosuke Imai: Statistical Analysis of List Experiments, Political Analysis, **20 (1)**, (2012), 47-77.

Blair, Graeme, Kosuke Imai, and Jason Lyall: *Comparing and Combining List and Endorsement Experiments: Evidence from Afghanistan*, American Journal of Political Science, **58 (4)**, (2014), 1043-1063.

Blair, Graeme, Kosuke Imai, Bethany Park, Alexander Coppock, and Winston Chou: *'Package list'*, available at `https://cran.r-project.org/web/packages/list/list.pdf`, last accessed on April 22, 2017.

Glynn, Adam: *What Can We Learn with Statistical Truth Serum?: Design and Analysis of the List Experiment*, Public Opinion Quarterly, **77**, (2013), 159-172.

Imai, Koisuke: *Multivariate Regression Analysis for the Item Count Technique*, Journal of the American Statistical Association, **106 (494)**, (2011), 407-416.

Krosnick, Jon A.: *Survey Research*, Annual Review of Psychology, **50**, (1999), 537-567.

Kramon, Eric, and Keith R. Weghorst: *Newsletter of the APSA Experimental Section*, **3 (2)**, (2012), 14-24.

Ostwald, Kai, and Guillem Riambau: *Voting Behavior under Doubts of Ballot Secrecy*, unpublished manuscript.

Rosenfeld, Bryn, Kosuke Imai, and Jacob Shapiro: *An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions*, American Journal of Political Science, **60 (3)**, (2016), 783-802.

Tsuchiya, Takahiro, Yoko Hirai and Shigeru Ono: *A Study of the Properties of the Item Count Technique*, The Public Opinion Quarterly, **71 (2)**, (2007), 253-272.

Wong, Maisy: *Estimating Ethnic Preferences Using Ethnic Housing Quotas in Singapore*, Review of Economic Studies, **80 (3)**, (2013), 1178-1214.