Who are our experts?

Predictors of participation in expert surveys

Anonymized authors

July 27, 2020

Abstract

Who are the colleagues participating when asked to complete expert surveys? This research note investigates which individuals' characteristics associate with positive responses. Drawing on an expert survey dedicated to post-conflict trials, we collect data on various attributes of both respondents and non-respondents such as their age, sex, academic positions, disciplines, and research outputs. We expect that decisions to participate result from an interplay of (1) individuals' levels of context-specific expertise, (2) the value attached to their expert role, (3) their confidence in making authoritative statements, and (4) resource constraints. Employing logistic regression models and statistical simulations (*N=414*), we find that context-specific expertise is the primary, but not the only determinant of participation. On the one hand and luckily, individuals whose research corresponds closely to the object of study are most likely to participate. On the other hand and unfortunately, individuals with high citation outputs, female experts, and Area Studies-scholars are less likely to respond. Consequently, certain groups are under-represented in expert evaluations frequently considered as authoritative source of knowledge.

1 Introduction

Who are the colleagues participating when asked to fill expert surveys? Expert surveys are a powerful research tool allowing the systematic study of hard-to-measure phenomena through the aggregation of specialized knowledge (Maestas 2016). While researchers often treat expert surveys as authoritative source of knowledge, little is known about self-selection dynamics underlying expert participation. The issue of non-response is well-studied in the context of survey research in general (e.g., Blom and Kreuter 2011; Groves and Peytcheva 2008); however, we contend that expert surveys deserve special attention. Expert surveys are unique in the type of contacted individuals and in the specific role designated as 'expert' offered to respondents.

Against this backdrop, we build on a recently administered expert survey on post-conflict trials to study individual-level determinants of participation decisions. We investigate why some contacted individuals decided to participate while others opted-out or simply ignored our survey invitation. Ideally, the willingness to participate in expert surveys would be entirely a function of the context-specific expertise of the contacted individuals. In other words, the more knowledge an individual commands about the object of research, the more likely s/he self-selects into the pool of respondents. However, there could be other systematic dynamics at play which could drive individuals' willingness to participate. For instance, it might be the case that different time capacities, diverging values attached to the expert role, or distinct levels of confidence drive self-selection of respondents. The key research question of our study is whether expert selection procedures select those with the most relevant expertise.

In order to investigate this question empirically, we compiled an original dataset on various attributes of all individuals contacted for our expert survey. We proxy context-specific expertise with the match between scholars' publications and the specific topic of the expert survey. Additionally, we collected data on ages, academic positions, academic disciplines, academic outputs, and locations of research. Drawing on logistic regression models and statistical simulations, we find that scholars whose academic work is closely related to the topic of the expert survey were more likely to participate. While context-specific expertise was the main driver of self-selection to our expert survey, we find that other factors led to systematic under-representations. Female scholars, scholars with high citations outputs, and Area Study scholars were less likely to participate. We explain these findings primarily with resource constraints that systematically affect the time capacities of certain scholars.

In the following section, we first describe our expert survey on post-conflict trials and provide a theoretical framework for participation decisions. Subsequently, we introduce our data

collection on attributes of the contacted individuals. Afterwards, we present our multivariate analyses providing evidence for systematic self-selection dynamics in expert survey participation. We conclude with a discussion of the generalisability and the limitations of our findings.

2 Background of the expert survey

2.1 An expert survey on post-conflict trials

The expert survey under scrutiny was substantively motivated by a systematic investigation of biases against the opposition in post-conflict trials. Post-conflict trials are judicial proceedings concerned with conflict-related violence implemented in five-years post-conflict windows (Binningsbø, Loyle, et al. 2012).[1] In order to generate comparable data on partiality levels of post-conflict trials, we conducted an expert survey evaluating all major post-conflict trials between 1946 and 2006 as recorded by the Post-Conflict Justice Dataset (Binningsbø, Loyle, et al. 2012). An overview of the survey items is presented in the Appendix. Experts were asked to rate the items on continuous Likert scales ranging from 0 to 10 in integers, whereby higher values capture higher degrees of procedural justice.

2.2 Expert selection procedure

We demanded fulfillment of the following three inclusion criteria to deem an individual as expert: 1.) An individual must have published a peer-reviewed article, a monograph, or a book chapter on the respective post-conflict trial under investigation or, alternatively, on the country where it was implemented. In the latter case, we demanded that the publication at least partially concerns the political context of the country during trial implementation. Solely publications from the fields of Anthropology, Area Studies, Economics, History, International Relations, Law, and Political Science were deemed as relevant. 2.) A scholar must hold at least the academic degree PhD to be considered as expert.[2] 3.) An individual must not be heavily biased towards any party involved in the trial. We screened experts' publications and biographies for reasonable degrees of political neutrality excluding politicians and co-ideologues of conflict parties.[3] Each scholar was only contacted about one specific post-conflict trial where s/he demonstrated particularly high levels of expertise.[4]

[1] Armed conflicts are defined as contested incompatibility concerning government and/or territory where the use of armed forces between two parties, of which at least one represents the government of a state, results in 25 battle-deaths within a year (Gleditsch et al. 2002: 618-619).

[2] Of course we are not claiming that only PhD-holders are experts, though, in order to err on the conservative and more academic side of expertise, we opted for this restrictive criterion.

[3] While we were generally lenient on this criterion, we solely excluded two potential experts for the International Criminal Tribunal for the former Yugoslavia that clearly promoted an ethnic agenda.

[4] A challenging point is that we do not know *a priori* the true population of experts. However, the pre-defined selection procedure makes explicit which types of experts are identified.

We performed the expert screening procedure with *Google Scholar*. The standard search string included the term 'trial' as well as the country and the year of trial implementation. If not sufficient publications could be found, we varied the search string adding specific names of conflict parties involved in the judicial proceedings. In so doing, we identified at least five experts for each of the 53 post-conflict trials under investigation.[5]

2.3 Survey administration

We used the online survey platform *Qualtrics* for survey administration. In total, we distributed 415 emails which contained short descriptions of the respective trial as provided by the Background Narratives of the Post-Conflict Justice Dataset (Binningsbø and Loyle 2012) and personalized invitations to the online survey. The invitation emails were administered through an Oxford University email account (a template is presented in the Online Appendix). Each contacted individual received two follow-up reminders if no response has been obtained within the course of two weeks. Experts were also provided with the option to participate anonymously given the sensitive nature of evaluations of post-conflict trials. The survey administration process was conducted between November 2017 and April 2018.

In total, 85 individuals completed the expert survey amounting to a response rate of 20.5%. Among the non-respondents, 41 individuals (9.9%) contacted us explaining that they lack sufficient in-depth expertise to rate the survey items in an adequate manner. The remaining 289 individuals (69.6%) did not react to the survey invitation. A detailed discussion of the survey and its substantive results can be found in [reference anonymised].

3 Systematic dynamics of expert self-selection

Who are the individuals that self-selected into the pool of respondents? And which types of individuals refused to answer the survey? We suggest that participation decisions result from an interplay of (1) experts' levels of context-specific expertise, (2) the value attached to the expert role, (3) their confidence in making authoritative statements, and (4) their respective resource constraints.

(1) Expert surveys strive for the inclusion of individuals with maximal degrees of expertise about the object under research. Consequently, the type of expertise required in expert surveys is highly context-specific being solely defined with regard to the particular object of study. Herein, context-specific expertise implies that a scholar's research specifically focuses on a post-

[5]We focused exclusively on major post-conflict trials identified by the process-scope variable in the Post-Conflict Justice Database.

conflict trial and its political context. We assume that context-specific expertise positively affects participation decisions since proficient individuals are likely to develop a certain desire to contribute to debates about their object of expertise. Further, individuals with in-depth knowledge have lower participation costs as they might be able to answer the questions in an *ad hoc* process. While the dimension of context-specific expertise would be ideally the only driver of expert self-selection, we argue that it is not the only one.

(2) We contend that the decision to participate is also affected by the value individuals attach to the expert role. The role of an expert is socially constructed relying on shared perceptions of competence. By contacting a scholar as an expert, s/he might feel appreciated and to some degree honored. However, the appreciation of the expert role might vary across respondents affecting participation decisions. For instance, the value scholars attach to their status as expert could be systematically affected by their positions in perceived academic hierarchies. It might be the case that scholars in earlier stages of their career value the expert role higher than renowned professors. The same might apply to scholars from smaller or non-Western universities and those with fewer citations. In contrast, senior professors from top universities might develop a certain level of response fatigue as result of being frequently contacted.

(3) We further suggest that participation decisions are systematically affected by individuals' confidence in making authoritative statements. Participation in expert surveys goes along with a certain responsibility as responses might shape commonly held perceptions about phenomena under investigation. This is particularly consequential in social contexts where widespread perceptions might indirectly affect patterns of behavior (e.g., Finnemore 1996). We suggest that individuals' level of confidence to make such statements could be systematically affected by their age and their gender (e.g., Kukulu et al. 2013; Orenstein 2013). Furthermore, we assume that training in different academic disciplines shapes attitudes towards the the complexity of social processes. As a result, we contend that scholars employing different methodological approaches differ in their confidence to rate social processes on ordinal scales.

(4) Finally, we argue that resource constraints affect the likelihood of expert participation. Available time resources to participate in voluntary tasks on top of the core academic obligations vary systematically across scholars. Previous research demonstrates that female scholars are structurally underrepresented in academia and systematically under-cited (e.g., Dion, Sumner, and Mitchell 2018; Maliniak, Powers, and Walter 2013). This disparity becomes even more acute in the case of parenthood since gendered child care responsibilities negatively affect academic productivity (Felisberti and Sear 2014; Hunter and Leahey 2010). These structural disadvantages suggest that female researchers have less time capacities to focus on additional

tasks. Furthermore, we suggest that resource constraints correlate with academic positions. Highly successful professors in senior positions could have too many other responsibilities to dedicate their time to expert surveys.

We assume that the quality of our survey responses depends on individuals' levels of expertise on the post-conflict trials under scrutiny. Hence, self-selection would be ideally only driven by context-specific expertise (dimension 1). We suggest that context-specific expertise is best captured by the specific fit of scholars' publications to the object of study. The closer the academic work of a scholar relates to the post-conflict trial under scrutiny, the higher the expected level of context-specific expertise. However, we hypothesize that also non-expertise related factors (dimensions 2-4) affect participation decisions. While we are unable to measure these dimensions directly, we use several observable attributes to identify determinants of self-selection that are unrelated to context-specific expertise.

4 Quantitative Analysis

4.1 Predictors

We collected data on the *year of birth* of the contacted individuals and on their sex captured by a binary indicator called *female*. We measure different stages of academic careers with measures for the *year of PhD* and experts' academic position differentiating between *professor* and other *post-PhD* positions. Further, we measure whether professors are already retired being *emeritus* professors. Experts' academic output is measured with their *number of publications* and with their *number of citations*. The location where experts conduct their research is captured by a measure that indicates *academic* positions and three measures that capture *Western institutions*, *US institutions*, and *Ivy League class* universities.

To account for experts' respective disciplines, we measure whether they are primarily trained in the fields of *Anthropology*, *Area Studies*, *Economics*, *History*, *Law*, *Sociology*, or *Political Science*. We further account for methodological differences by capturing whether an expert works primarily as *quantitative scholar*. Additionally, we measure the *year of publication* on which selection was based to capture potential oblivion or recency effects.

Finally, to capture whether self-selection is a function of context-specific in-depth knowledge, we evaluate the *match* of the selection publication with the post-conflict trial under investigation. This ordinal variable was coded as 2, if the publication is directly dedicated to the trial under scrutiny. It was coded as 1, if the publication deals primarily with the political situation in the country during trial implementation and as 0, if the publication is only loosely connected to the

trial and its political context. A more detailed overview of all explanatory variables and their respective operationalizations is presented in the Appendix.

4.2 Outcome variables

To study the impact of these predictors on response decisions, we created two different outcome variables. The first outcome variable called *responded* is a binary measure of responses to our expert survey. If experts participated at least until the half of the survey, we coded the indicator variable as 1. If experts did not respond or abandoned the survey before the half, we coded the variable as 0. In total, 85 individuals (20.5%) responded to the expert survey.

The second outcome variable termed *excused* is a binary measure capturing whether contacted experts replied to us justifying non-response with a lack of competence in the specific field of research. Their excuse took generally the form of an informal email where individuals claimed that they are unable to evaluate the respective questions in an adequate manner. The variable is coded as 1 for individuals that issued such excuses and as 0 for both experts who completed the survey and for those who did not respond at all. In total, 41 individuals (9.9%) excused their non-participation with a lack of context-specific in-depth expertise. Descriptive statistics for all variables are presented in the Appendix.

4.3 Statistical models

We run logistic regression models with robust standard errors to account for the binary nature of our outcome variables. Given the substantial imbalance between positive and negative values on the binary dependent variables, we additionally run logistic rare events models for each model specification (*see* King and Zeng 2001). Several of our predictors are highly correlated such as *# of citations* and *# of publications*. To avoid multicollinearity, we included these predictors into separate models if Variance Inflation Factors were overly high (alternative model specifications are presented in the Appendix). The variables *year of birth* and *year of PhD* display a substantial number of missing values which led us to exclude them from the main models.[6] As a robustness test, we imputed missing values for these variables following the approach of Royston et al. (2009) and added them to the main model specifications. Further, we run multinomial logistic regression models to test whether the findings hold if completing and excusing are not treated as independent.[7] We also tested for theoretically relevant interactions but could not trace any significant effects.

[6] If we could not trace this information online, we contacted scholars asking them directly for these years. However, several scholars refused to respond likely to due the sensitive character of age-related information.

[7] In the multinomial logistic regression, we code the outcome variable as 2 for survey completions, as 1, for excuses, and as 0, for non-responses.

To derive substantively meaningful quantities of interest, we subsequently simulated predicted probabilities by drawing 10,000 simulations of the parameters from our logistic regression models (King, Tomz, and Wittenberg 2000: 348). Building on the approximated probability distributions, we estimated the individual effect of all substantively influential predictors holding continuous variables at their mean and binary variables at their mode.

5 Results of the multivariate analyses

Our two logistic regression models on the outcome variables *completed* (Model 1) and *excused* (Model 2) are presented in Table 1.[8] Based on these logistic regression models, we ran simulations of the statistically significant effects.[9]

For Model 1, the predicted likelihood of participation is 49% if there is a close *match* between a scholars' publication and the object of study (95% confidence interval: 12%, 87%). If the *match*-variable was coded in its mid-category indicating that a publication deals not primarily with the object of study, the predicted likelihood of participation is 29% (95% confidence interval: 5%, 67%). Finally, if the *match*-variable was coded in its lowest category signifying an incidental relation to the object of study, the predicted probability of participating is 13% (95% confidence interval: 2%, 42%). Thus, the likelihood of obtaining a response from a scholar with a loosely matching publication is 36% lower than from a scholar with a closely fitting publication (95% confidence interval: 10%, 59%). To find out whether first differences on the *match*-scale are substantively meaningful, we simulated them 10,000 times. Indeed, for each one-step increase on the *match*-variable 10,000 values were obtained higher than 0. The simulated predicted values and their 95% confidence intervals are illustrated in Figure 1.[10]

[8] We omit the predictor *sociology* from the second model as no sociologist excused non-participation.

[9] We also re-run the models with standardized effects. We present these models in the Online Appendix that can be accessed on the homepage of the first author.

[10] We acknowledge that the explicit reference to the respective publication in invitation emails may prime experts to think carefully through the match with the topic. This could potentially exacerbate the effect of perceived competence on positive response.

Table 1: Logistic regression models

	(1) Completed	(2) Excused
Female	-0.735*	0.162
	(-2.04)	(0.40)
Professor	0.133	0.752
	(0.40)	(1.50)
Academic	-0.039	-1.082
	(-0.07)	(-1.90)
Log Citations	-0.200*	0.127
	(-2.08)	(0.85)
Western	-0.465	0.853
	(-1.13)	(1.26)
US	0.463	0.0468
	(1.43)	(0.12)
Quantitative scholar	-0.796	0.285
	(-1.59)	(0.47)
Match of publication	1.084***	-0.623*
	(5.02)	(-2.20)
Emeritus	0.446	0.258
	(1.09)	(0.49)
Ivy League class	-0.460	0.627
	(-0.73)	(1.12)
Year of publication	0.031	0.001
	(1.58)	(0.04)
Political Science	0.594	1.712
	(0.74)	(1.66)
Anthropology	1.187	0.215
	(1.38)	(0.15)
Area Studies	-1.340	2.596*
	(-1.00)	(2.32)
Economics	1.151	2.226
	(1.03)	(1.66)
History	0.355	1.486
	(0.44)	(1.36)
Law	-1.166	1.243
	(-1.02)	(0.86)
Sociology	-0.142	
	(-0.14)	
Constant	-62.72	-6.732
	(-1.59)	(-0.15)
Observations	404	404

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

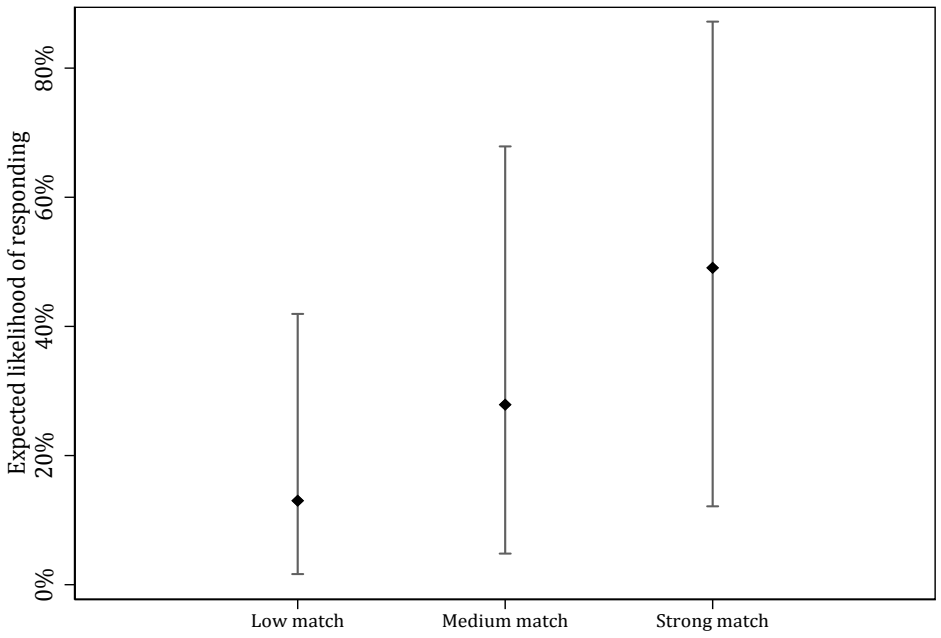

Figure 1: Statistical simulation match-variable (based on model 1)

Considering different *citation* outputs, we compared predicted probabilities between the 25th and the 75th percentile of the continuous logged *citations*-variable. The predicted likelihood of responding is 15% when the logged *citations*-variable is in its 25th percentile (95% confidence interval: 2%, 42%) and drops to 11% when the logged *citations*-variable is in its 75th percentile (95% confidence interval: 1%, 37%). We simulated the difference between the 25th and the 75th percentile 10,000 times and obtained 9,787 values higher than 0.[11] With regards to respondents' sex, the predicted likelihood of survey participation is 7% (95% confidence interval: 1%, 28%) for *females* and 13% (95% confidence interval: 2%, 42%) for males. Running 10,000 simulations of the first difference, 9,789 parameters were obtained greater than 0. Thus, the first difference is higher than 0 with a probability of 98%.

For Model 2, the effect of the *match*-variable operates in the opposite direction. The predicted likelihood of excuses is 2% for scholars with closely matching publications (95% confidence interval: 1%, 11%) and 7% for scholars with loosely matching publications (95% confidence interval: 1%, 28%). We find a more pronounced effect of scholars' academic training in *Area Studies*. Scholars whose main discipline was identified as *Area Studies* excuse non-participation with a likelihood of 39% (95% confidence interval: 13%, 71%). In contrast, scholars from the reference group containing all other academic disciplines opt-out with a predicted likelihood of 7% (95% confidence interval: 1%, 28%). Simulating this first difference 10,000 times, we

[11] The effect of the *logged citations*-variable should be taken with a grain of salt as this predictor loses significance in a few alternative specifications (as demonstrated in the Appendix).

obtained 9,906 values larger than 0. This implies that the first difference exceeds 0 with a probability of 99%. The predicted probabilities for the *Area Studies*-variable are illustrated in Figure 2. The results remain robust to the imputation of missing values for *year of birth* and *year of PhD*. Nevertheless, the findings should be taken with a grain of salt due to the comparatively low statistical power of the analysis.

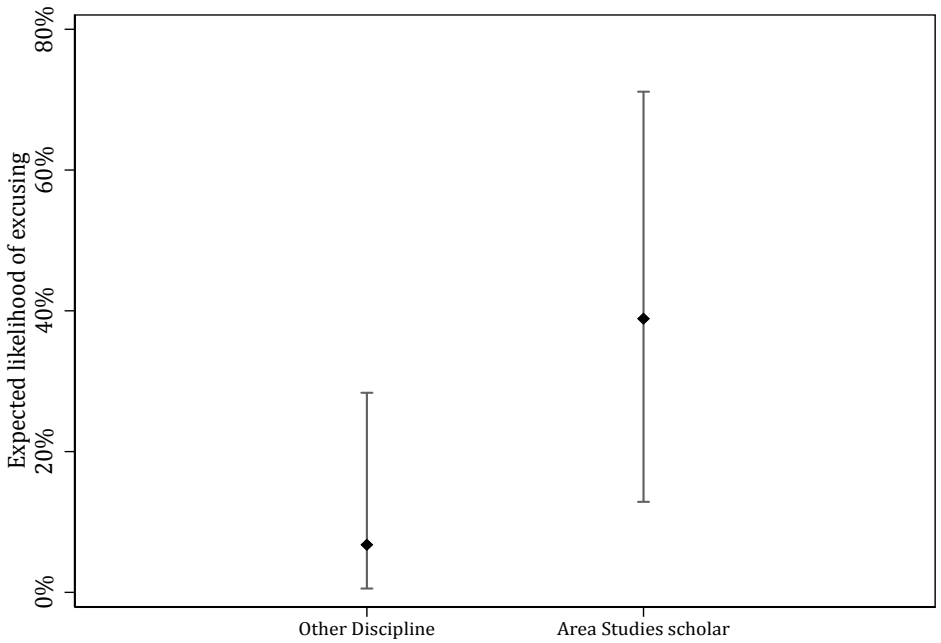

Figure 2: Statistical simulation Area Studies-variable (based on model 2)

6 Discussion

Does self-selection automatically filter 'true' experts from 'ostensible' experts operating along the dimension of context-specific knowledge? Against the backdrop of these findings, there emerges a mixed answer to this question. Indeed, the key factor driving individuals' decisions to participate in our survey is the match between their research focus and the object of study. This implies that high levels of in-depth knowledge about specific topics increase individuals' willingness to participate in surveys dedicated to these topics. Further, it suggests that individuals lacking such context-specific expertise tend to opt-out providing an expedient safeguard against poor survey responses.

However, this possible positive self-selection process co-exists with other systematic correlations underlying individual response decisions. Scholars with high citation outputs were less likely to participate even when controlling for their respective level of context-specific expertise. It is most likely that citation outputs account for academic success - or at least for commonly

held perceptions about it - implying that renowned scholars are less likely to answer. Given that successful scholars are most likely to be identified and therefore frequently contacted for suchlike research projects, it is presumable that they develop a certain level of response fatigue. Thus, our expert survey might under-represent the most prominent voices in the research field creating needs for selective incentives to elicit their responses.

Furthermore, this study finds that female scholars are less likely to participate than male colleagues.[12] We can only speculate about the reasons for this under-representation in absence of further information about underlying causal mechanisms. Previous research suggests that socialization-related factors create a 'confidence gap' between men and women (e.g., Orenstein 2013; Sax and Harper 2007). Mirroring this argument, it might be the case that female scholars are more hesitant to claim the expert role or more willing to acknowledge competence deficits than male scholars. Of course, a further explanation is that in academia female colleagues are structurally underrepresented - the infamous "glass ceiling" -, systematically under-cited (e.g., Dion, Sumner, and Mitchell 2018; Maliniak, Powers, and Walter 2013) and, therefore, they have all the incentive structure to keep focusing on their own research.[13]

Finally, we find that scholars from the field of *Area Studies* are more likely to opt-out by explicitly excusing their non-participation. This finding might reflect different perceptions about the inherent quantifiability of social processes. While scholars from disciplines such as Political Science, Sociology, or Economics are used to classify social contexts on ordinal scales, different methodological and epistemological conceptions might create reservations about such classifications. We presume that these dynamics drive disproportional numbers of excuses from Area Studies-scholars. Indeed, several justifications for non-participation explicitly referred to doubts about the meaningfulness of comparative large-N research in the context of multi-dimensional social phenomena such as 'trial fairness'. However, the finding should be carefully interpreted in light of the comparatively low number of Area Study scholars in our sample.

No evidence could be found that non-expertise related dynamics lead to heightened response propensities. In contrast, the delineated effects all capture diminished response likelihoods for certain subgroups. This suggests that non-expertise related dynamics make certain scholars less likely to participate creating selective under-representations. However, it would be far more consequential if such dynamics would increase response likelihoods among certain subgroups

[12]This finding speaks to previous research suggesting that females do disproportional amounts of service (Guarino and Borden 2017). However, the male-female differential is largely driven by internal service (service to the university, campus, or department) instead of external service (service to the university, campus, or department). Arguably, participation in expert surveys falls in the category of external service by enhancing external recognition of experts.

[13]The finding is noteworthy given that females tend to be over-represented in non-expert surveys (e.g., Cheung et al. 2017; Volken 2013).

implying that scholars with limited context-specific expertise tend to self-select in the pool of respondents. This reassuring finding aligns with the results of a recent study on the correlates of expert reliability in the V-Dem project finding little evidence of theoretically-untenable bias due to expert characteristics (Marquardt et al. 2018).

7 Conclusion

To summarize, the positive news is that our experts seem actually to be experts that have been writing and researching about the specific study objects. However, these colleagues who agreed to participate tend to systematically under-represent female scholars and Area Studies-scholars. Therefore, our expert survey could be skewed against important contributions from larger and more diverse scholarly perspectives.

The question arises to what extent the dynamics investigated in this expert survey are generalizable to other expert surveys employed for various study objects in diverse academic fields. Our expert survey is somewhat idiosyncratic in its focus on a normatively loaded and inherently multi-dimensional concept such as 'trial fairness' which could impact response hesitance. Indeed, our response rate of 20.5% is below average. By retrieving the response rates from various major Political Science studies drawing on online expert surveys, we identified that the average response rate was 37.4% (sd: 13.8).[14] Hence, our findings might only generalize to expert surveys on study objects that are similarly sensitive and multi-dimensional. More research is necessary to confirm these self-selection processes in other types of expert surveys.

We also acknowledge that our analysis is unable to disentangle several of our suggested theoretical dimensions. For instance, the negative effect of female-measure could operate through the pathway of resource constraints or through the pathway of confidence. Since we cannot directly measure these dimensions, we can only speculate about the causal mechanisms. In essence, our analysis shows that factors that are not genuinely related to expertise affect response propensities. However, we cannot empirically establish through which mechanisms these factors affect self-selection.

Further, we acknowledge that our expert selection procedure was comparatively restrictive by demanding a PhD and reasonable degrees of political neutrality to consider an individual as experts. Generally, there is a trade-off between the quality of experts and the size of the

[14] We used the search terms 'expert survey' and 'politics' in Google Scholar and collected the response rates of all articles identified on the first ten search pages. We retrieved the response rates of 29 studies with an average response rate of 42.8% (sd: 18.0). For the subset of online expert surveys, the average response rate is 37.4% (12 studies). All the identified studies with their respective response rates are presented in Table 8 in the Appendix.

respondent pool. While only few individuals were filtered through these criteria, our selection procedure tends to prioritize the quality of the experts. This might lead to above average values on the *match*-variable compared to other expert surveys. In light of these caveats and the comparatively low statistical power of our analysis, the findings of this study should be cautiously interpreted as tentative evidence. Given that this is the first systematic empirical investigation of expert self-selection dynamics, we hope to encourage similar investigations in different fields that could complement our findings.

This short piece calls for more research on systematic measurement errors (e.g., Ruggeri, Gizelis, and Dorussen 2011) and awareness of possible shortcomings of what we tend to shield with the 'expert' label. Finally, academia should indeed rethink incentive structures to give the opportunity to all types of scholars to contribute to our data generation processes thanks to their knowledge.

References

Arvanitidis, Paschalis A., George Petrakos, and Sotiris Pavleas (2010). "On the dynamics of growth performance: an expert survey". In: *Contributions to Political Economy* 29.1, pp. 59–86.

Azzi, Stephen and Norman Hillmer (2013). "Evaluating Prime Ministerial Leadership in Canada: The Results of an Expert Survey". In: *Canadian Political Science Review* 7.1, pp. 13–23.

Bakker, Ryan et al. (2015). "Measuring party positions in Europe: The Chapel Hill expert survey trend file, 1999–2010". In: *Party Politics* 21.1, pp. 143–152.

Binningsbø, Helga and Cyanne Loyle (2012). *Armed Conflict and Post-Conflict Justice Dataset: Background Narratives*. Ed. by PRIO: Centre for the Study of Civil War. URL: `http://www.justice-data.com/pcj-dataset/`.

Binningsbø, Helga, Cyanne Loyle, Scott Gates, and Jon Elster (2012). "Armed conflict and post-conflict justice, 1946–2006: A dataset". In: *Journal of Peace Research* 49.5, pp. 731–740.

Blom, Annelies G. and Frauke Kreuter (2011). "Special issue on survey nonresponse". In: *Journal of Official Statistics: JOS* 27.2.

Bowler, Shaun, David M. Farrell, and Robin T. Pettitt (2005). "Expert opinion on electoral systems: So which electoral system is "best"?" In: *Journal of Elections, Public Opinion & Parties* 15.1, pp. 3–19.

Castles, Francis G. and Peter Mair (1984). "Left–right political scales: Some 'expert' judgments". In: *European Journal of Political Research* 12.1, pp. 73–88.

Chernykh, Svitlana, David Doyle, and Timothy J. Power (2017). "Measuring Legislative Power: An Expert Reweighting of the Fish–Kroenig Parliamentary Powers Index". In: *Legislative Studies Quarterly* 42.2, pp. 295–320.

Cheung, Kei Long, M. Peter, Cees Smit, Hein de Vries, and Marcel E. Pieterse (2017). "The impact of non-response bias due to sampling in public health studies: a comparison of voluntary versus mandatory recruitment in a Dutch national survey on adolescent health". In: *BMC public health* 17.1, pp. 276–286.

Coma, Feerran Martinez I. and Carolien van Ham (2015). "Can experts judge elections? Testing the validity of expert judgments for measuring election integrity". In: *European Journal of Political Research* 54.2, pp. 305–325.

Dahlberg, Stefan, Carl Dahlström, Petrus Sundin, and J. Teorell (2013). "The Quality of Government Expert Survey 2008–2011: A Report". In: *QoG Working Paper Series* 15.

David, Roman and Ian Holliday (2012). "International sanctions or international justice? Shaping political development in Myanmar". In: *Australian Journal of International Affairs* 66.2, pp. 121–138.

Dion, Michelle L., Jane Lawrence Sumner, and Sara McLaughlin Mitchell (2018). "Gendered citation patterns across political science and social science methodology fields". In: *Political Analysis* 26.3, pp. 312–327.

Felisberti, Fatima M. and Rebecca Sear (2014). "Postdoctoral researchers in the UK: a snapshot at factors affecting their research output". In: *PloS one* 9.4.

Finnemore, Martha (1996). "Norms, culture, and world politics: insights from sociology's institutionalism". In: *International organization* 50.2, pp. 325–347.

Gervasoni, Carlos (2010). "Measuring variance in subnational regimes: Results from an expert-based operationalization of democracy in the Argentine provinces". In: *Journal of Politics in Latin America* 2.2, pp. 13–52.

Gleditsch, Nils Petter, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg, and Håvard Strand (2002). "Armed conflict 1946-2001: A new dataset". In: *Journal of Peace Research* 39.5, pp. 615–637.

Groves, Robert M. and Emilia Peytcheva (2008). "The impact of nonresponse rates on nonresponse bias: a meta-analysis". In: *Public opinion quarterly* 72.2, pp. 167–189.

Guarino, Cassandra M. and Victor M. H. Borden (2017). "Faculty service loads and gender: Are women taking care of the academic family?" In: *Research in Higher Education* 58.6, pp. 672–694.

Huber, John and Ronald Inglehart (1995). "Expert interpretations of party space and party locations in 42 societies". In: *Party Politics* 1.1, pp. 73–111.

Hunter, Laura A. and Erin Leahey (2010). "Parenting and research productivity: New evidence and methods". In: *Social Studies of Science* 40.3, pp. 433–451.

Kato, Junko and Michael Laver (1998). "Party policy and cabinet portfolios in Japan, 1996". In: *Party Politics* 4.2, pp. 253–260.

— (2003). "Policy and Party Competition in Japan after the Election of 2000". In: *Japanese Journal of Political Science* 4.1, pp. 121–133.

Kerby, Matthew and Kelly Blidook (2014). "Party policy positions in Newfoundland and Labrador: Expert survey results in the buildup to the 2011 provincial election". In: *American Review of Canadian Studies* 44.4, pp. 400–414.

King, Gary, Michael Tomz, and Jason Wittenberg (2000). "Making the most of statistical analyses: Improving interpretation and presentation". In: *American Journal of Political Science* 44.2, pp. 341–355.

King, Gary and Langche Zeng (2001). "Logistic regression in rare events data". In: *Political Analysis* 9.2, pp. 137–163.

Kukulu, K., O. Korukcu, Y. Ozdemir, A. Bezci, and C. Calik (2013). "Self–confidence, gender and academic achievement of undergraduate nursing students". In: *Journal of psychiatric and mental health nursing* 20.4, pp. 330–335.

Laver, Michael (1998). "Party policy in Ireland 1997 results from an expert survey". In: *Irish political studies* 13.1, pp. 159–171.

Lupu, Noam and Kristin Michelitch (2018). "Advances in survey methods for the developing world". In: *Annual Review of Political Science* 21, pp. 195–214.

Maestas, Cherie (2016). "Expert surveys as a measurement tool - challenges and new frontiers". In: *The Oxford Handbook of Polling and Survey Methods*. Ed. by Atkeson, Lonna Rae & Michael Alvarez. Oxford: Oxford Handbooks Online. URL: https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780190213299.001.0001/.

Maliniak, Daniel, Ryan Powers, and Barbara F. Walter (2013). "The gender citation gap in international relations". In: *International organization* 67.4, pp. 889–922.

Marquardt, Kyle L., Daniel Pemstein, Brigitte Seim, and Yi-ting Wang (2018). "What Makes Experts Reliable?" In: *V-Dem Working Paper* 68. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3190946.

McElroy, Gail and Kenneth Benoit (2007). "Party groups and policy positions in the European Parliament". In: *Party Politics* 13.1, pp. 5–28.

McLean, Iain, André Blais, James C. Garand, and Micheal Giles (2009). "Comparative journal ratings: A survey report". In: *Political Studies Review* 7.1, pp. 18–38.

O'Malley, Eoin (2007). "The power of prime ministers: Results of an expert survey". In: *International Political Science Review* 28.1, pp. 7–27.

Orenstein, Peggy (2013). *Schoolgirls: Young women, self esteem, and the confidence gap*. New York City: Anchor.

Pétry, François, Benoît Collette, and Hans-Dieter Klingemann, eds. (2012). *Left-Right in Canada: Comparing data from party manifesto content and expert surveys*.

Polk, Jonathan et al. (2017). "Explaining the salience of anti-elitism and reducing political corruption for political parties in Europe with the 2014 Chapel Hill Expert Survey data". In: *Research & Politics* 4, pp. 1–9.

Ray, Leonard (1999). "Measuring party orientations towards European integration: Results from an expert survey". In: *European Journal of Political Research* 36.2, pp. 283–306.

Ray, Leonard and Hanne Marthe Narud (2000). "Mapping the Norwegian political space: Some findings from an expert survey". In: *Party Politics* 6.2, pp. 225–239.

Rohrschneider, Robert and Stephen Whitefield (2007). "Representation in new democracies: Party stances on European integration in post-communist Eastern Europe". In: *The Journal of Politics* 69.4, pp. 1133–1146.

Royston, Patrick, John B. Carlin, and Ian R. White (2009). "Multiple imputation of missing values: new features for mim". In: *The Stata Journal* 9.2, pp. 252–264.

Ruggeri, Andrea, Theodora-Ismene Gizelis, and Han Dorussen (2011). "Events data as Bismarck's sausages? Intercoder reliability, coders' selection, and data quality". In: *International Interactions* 37.3, pp. 340–361.

Sax, Linda J. and Casandra E. Harper (2007). "Origins of the gender gap: Pre-college and college influences on differences between men and women". In: *Research in Higher Education* 48.6, pp. 669–694.

Schmitt, Hermann and Thomas Loughran, eds. (2017). *Understanding Ideological Change in Britain: Corbyn, BREXIT, and the BES Expert Surveys.*

Steinert, Christoph (2019). "Trial fairness before impact: Tracing the link between post-conflict trials and peace stability". In: *International Interactions*. URL: https://doi.org/10.1080/03050629.2019.1657114.

Szöcsik, Edina and Christina Isabel Zuber (2015). "EPAC–a new dataset on ethnonationalism in party competition in 22 European democracies". In: *Party Politics* 21.1, pp. 153–160.

Volken, Thomas (2013). "Second-stage non-response in the Swiss health survey: determinants and bias in outcomes". In: *BMC public health* 13.1, pp. 167–177.

Warwick, Paul (2005). "Do policy horizons structure the formation of parliamentary governments?: The evidence from an expert survey". In: *American Journal of Political Science* 49.2, pp. 373–387.

Appendices

A Overview of Predictors

Personal characteristics:

Female: Binary indicator of the sex of the contacted expert.

Year of birth: Continuous indicator of the year of birth of the contacted expert. If available, we used information in CVs and on homepages of scholars. Otherwise, we contacted them asking for this information.

Stages of academic careers:

Year of PhD: Continuous indicator capturing the year when an expert completed her/his PhD.

Professor: Binary measure capturing whether the contacted expert holds a full professorship.

Post-PhD: Binary variable indicating whether an expert holds a position as post-Doc, Assistant/ Associate/ Junior Professor. Coded as 0, if an expert holds a full professorship.

Emeritus: Binary indicator recording whether an expert is a retired professor.

Academic output:

Number of publications: Ordinal measure of the number of publications. If an expert has 1-5 publications, coded as 0. If s/he has 6-15 publications, coded as 1. If s/he has more than 15 publications, coded as 2.

Number of citations: Continuous variable indicating the number of citations in academic journals. We used Google Scholar profiles to collect this data. If no Google Scholar profiles were available, we added up the number of citations of authors' identifiable publications ourselves using references in Google Scholar.

Location of research:

Academic: Binary measure indicating whether an expert works currently at a university. If s/he conducts research at a non-academic institution or s/he has a non-research job, coded as 0.

Western institution: Binary measure indicating whether an expert works at a US or European university or institution.

US institution: Binary measure capturing whether an expert works at a US institution.

Ivy League class university: Binary variable capturing whether an expert is employed at a 'self-declared' top university. The following universities are deemed as Ivy League class universities: US Ivy League, Oxford, Cambridge, Science Po.

Academic discipline/ Research approach:

Anthropology: Binary indicator capturing whether an expert is primarily trained as anthropologist.

Area Studies: Binary measure signifying whether an expert is primarily trained as a specialist for area studies.

Economics: Binary variable capturing whether an expert is primarily trained as economist.

History: Binary measure indicating whether an expert is primarily trained as historian.

Law: Binary indicator denoting whether an expert is primarily trained as a lawyer.

Political Science: Binary variable indicating whether an expert is primarily trained in the field of political science.

Quantitative scholar: Binary measure capturing whether an expert works primarily with quantitative methods.

Specific expertise for object of research:

Match of publication: Ordinal variable capturing whether the selection publication pertains directly to the post-conflict trial. We hand-coded this variable reading Abstracts and screening full texts of scholars' publications. Coded as 2, if the selection publication contains the respective post-conflict trial already in its title or abstract. Coded as 1, if the does not address the trial in its title or abstract but pertains directly to the political situation in the country during trial implementation. If the selection publication is only loosely connected to the post-conflict trial, coded as 0.

B Tables

Table 2: The Post-Conflict Trials Expert Survey

# of Item	Item	Extreme poles of scale (continuous in integers)
1	Were all perpetrators of violence treated in an equal way or were some groups systematically discriminated?	0 = Unequal treatment 10 = Equal treatment
2	Were there indications that the government justified repression with reference to the justice process?	0 = Occurred frequently 10 = Never occurred
3	Were there incidences of violence related to the justice process such as targeting of judges and witnesses or retribution violence directed at perpetrators?	0 = Widespread violence 10 = Absence of violence
4	Did the scope of the process mandate concern only human rights violations perpetrated by certain groups or was violence from all sides (including the current government) considered?	0 = Extremely narrow focus 10 = Complete inclusiveness
5	Was the justice process restricted to a singular event or period of time or did it also concern potential backlash violence after the conflict/ genocide under investigation?	0 = Timewise restricted 10 = Timewise unrestricted
6	Did the narrative created by the justice process serve the purpose to consolidate the government?	0 = Distorted narrative 10 = Objective narrative
7	On a continuum from 0 to 10, whereby 10 indicates post-conflict fairness and 0 indicates post-conflict **in**justice: How would you evaluate the respective justice process overall?	0 = Post-conflict **in**justice 10 = Post-conflict justice

Table 3: Descriptive Statistics

VARIABLES	N	mean	sd	min	max	Completed	Excused
Female	414	0.273	0.446	0	1	15%	8%
Professor	414	0.486	0.500	0	1	19%	12%
Post-PhD	414	0.396	0.490	0	1	24%	5%
Western	414	0.845	0.362	0	1	19%	11%
US	414	0.539	0.499	0	1	22%	10%
Ivy League class	414	0.082	0.275	0	1	12%	18%
Academic	414	0.877	0.329	0	1	20%	9%
Quantitative	414	0.123	0.329	0	1	14%	16%
Emeritus	414	0.114	0.318	0	1	23%	15%
Political Science	414	0.457	0.499	0	1	26%	10%
Anthropology	414	0.085	0.279	0	1	31%	3%
Area Studies	414	0.070	0.256	0	1	3%	24%
Economics	414	0.029	0.168	0	1	17%	25%
History	414	0.213	0.410	0	1	17%	10%
Law	414	0.048	0.215	0	1	10%	5%
Sociology	414	0.051	0.220	0	1	14%	0%
# of publications	414	1.739	0.530	0	2	20% (at 2)	10% (at 2)
Match of publication	414	0.529	0.698	0	2	25% (at 2)	0% (at 2)
Year of publication	414	2006	8.376	1972	2017	-	-
Log # of citations	404	5.892	1.515	0	9.691	-	-
Year of birth	222	1956	12.91	1926	1991	-	-
Year of PhD	220	1992	14.23	1952	2017	-	-

Table 4: Summary results statistical simulations

	Low bound (95% CI)	Pr. mean	Up bound (95% CI)
Model 1: Completed			
High Match (2)	0.1214	0.4907	0.8720
Med. Match (1)	0.0482	0.2788	0.6788
Low Match (0)	0.0164	0.1301	0.4194
Log Citations (p25)	0.0190	0.1484	0.4675
Log Citations (p75)	0.0137	0.1113	0.3690
Female	0.0067	0.0743	0.2753
Male	0.0164	0.1301	0.4194
Model 2: Excused			
High Match (2)	0.0014	0.0228	0.1082
Med. Match (1)	0.0030	0.0386	0.1712
Low Match (0)	0.0054	0.0676	0.2836
Area Studies	0.1286	0.3888	0.7113
Other Discipline	0.0054	0.0676	0.2836

Table 5: Alternative model specifications (regressed on 'Completed')

	(1)	(2)	(3)	(4)	(5)	(6)
Female	-0.724*	-0.759*	-0.712*	-0.712*	-0.717*	-0.724*
	(-2.02)	(-2.15)	(-1.96)	(-1.97)	(-1.98)	(-2.00)
Western	-0.480	-0.662	-0.179	-0.179	-0.192	-0.190
	(-1.18)	(-1.71)	(-0.49)	(-0.49)	(-0.53)	(-0.52)
Quantitative Scholar	-0.824	-0.934	-0.716	-0.716	-0.713	-0.700
	(-1.62)	(-1.87)	(-1.44)	(-1.43)	(-1.44)	(-1.41)
Match of publication	1.080***	1.151***	1.079***	1.079***	1.083***	1.095***
	(5.05)	(5.51)	(5.04)	(5.04)	(5.03)	(5.09)
Year of publication	0.0291	0.0367	0.0286	0.0286	0.0286	0.0255
	(1.48)	(1.81)	(1.47)	(1.47)	(1.46)	(1.35)
Political Science	0.664	0.795	0.586	0.586	0.626	0.602
	(1.18)	(1.40)	(1.04)	(1.04)	(1.11)	(1.08)
Anthropology	1.263	1.485*	1.172	1.172	1.213	1.241
	(1.92)	(2.29)	(1.80)	(1.81)	(1.86)	(1.90)
Area Studies	-1.229	-1.089	-1.505	-1.506	-1.489	-1.456
	(-0.99)	(-0.88)	(-1.23)	(-1.24)	(-1.25)	(-1.22)
Economics	1.245	1.292	1.113	1.112	1.160	1.100
	(1.29)	(1.33)	(1.14)	(1.15)	(1.20)	(1.14)
History	0.454	0.515	0.412	0.412	0.455	0.495
	(0.77)	(0.87)	(0.70)	(0.70)	(0.77)	(0.84)
Law	-1.085	-1.095	-1.094	-1.094	-1.047	-1.125
	(-1.10)	(-1.12)	(-1.12)	(-1.12)	(-1.07)	(-1.17)
Emeritus	0.488	0.391	0.480	0.479	0.464	
	(1.22)	(0.95)	(1.18)	(1.19)	(1.14)	
Ivy League Class	-0.495	-0.570	-0.404	-0.402		
	(-0.80)	(-0.87)	(-0.65)	(-0.65)		
Academic	-0.0246	0.00570	0.0102			
	(-0.05)	(0.01)	(0.02)			
US	0.490	0.566				
	(1.51)	(1.77)				
Post-PhD	0.119					
	(0.40)					
Log citations	-0.174		-0.212*	-0.212*	-0.217*	-0.209*
	(-1.90)		(-2.21)	(-2.22)	(-2.27)	(-2.18)
Professor		-0.0867	0.184	0.186	0.205	0.250
		(-0.27)	(0.55)	(0.57)	(0.63)	(0.78)
Num. of publications		-0.116				
		(-0.39)				
Observations	404	414	404	404	404	404

t statistics in parentheses

Table 6: Rare events logistic regression models

	(1) Completed	(2) Excused
Female	-0.676*	0.178
	(-1.97)	(0.46)
Professor	0.124	0.697
	(0.39)	(1.46)
Academic	-0.0638	-1.032
	(-0.13)	(-1.90)
Log citations	-0.188*	0.112
	(-2.06)	(0.79)
Western	-0.448	0.700
	(-1.14)	(1.08)
US	0.429	0.0274
	(1.39)	(0.07)
Quantitative scholar	-0.690	0.327
	(-1.45)	(0.57)
Match of publication	1.022***	-0.553*
	(4.95)	(-2.04)
Emeritus	0.441	0.279
	(1.13)	(0.55)
Ivy League Class	-0.344	0.646
	(-0.58)	(1.20)
Year of publication	0.0279	-0.0004
	(1.50)	(-0.02)
Political Science	0.377	1.210
	(0.49)	(1.22)
Anthropology	0.955	0.200
	(1.16)	(0.14)
Area Studies	-1.049	2.066
	(-0.82)	(1.92)
Economics	1.005	1.710
	(0.94)	(1.33)
History	0.153	1.008
	(0.20)	(0.97)
Law	-1.117	1.158
	(-1.03)	(0.84)
Sociology	-0.219	
	(-0.23)	
Constant	-56.70	-3.086
	(-1.51)	(-0.07)
Observations	404	404

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7: Multinomial logistic regression

	(1) Excused	(2) Completed
Female	0.0525	-0.729*
	(0.13)	(-2.01)
Professor	0.763	0.193
	(1.52)	(0.57)
Academic	-1.102	-0.151
	(-1.90)	(-0.28)
Citations_log	0.0992	-0.189
	(0.66)	(-1.94)
Western	0.789	-0.407
	(1.16)	(-0.99)
US	0.127	0.479
	(0.32)	(1.46)
Quantitative Scholar	0.143	-0.781
	(0.24)	(-1.55)
Match of publication	-0.385	1.039***
	(-1.26)	(4.74)
Emeritus	0.341	0.499
	(0.63)	(1.22)
Elite university	0.595	-0.386
	(1.06)	(-0.61)
Publication year	0.00635	0.0319
	(0.28)	(1.61)
Political Science	1.837	0.800
	(1.78)	(1.44)
Anthropology	0.425	1.279*
	(0.29)	(1.98)
Area Studies	2.504*	-1.022
	(2.23)	(-0.82)
Economics	2.446	1.490
	(1.82)	(1.55)
History	1.551	0.525
	(1.42)	(0.90)
Law	1.037	-1.013
	(0.71)	(-1.04)
Constant	-17.20	-65.07
	(-0.38)	(-1.63)
Observations	404	404

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8: Response rates across experts surveys in Political Science

Publication	**Response Rate**	**Administered Online**
Arvanitidis, Petrakos, and Pavleas 2010	73%	×
Azzi and Hillmer 2013	57.4%	✓
Bakker et al. 2015	34.9%	×
Bowler, Farrell, and Pettitt 2005	31%	×
Castles and Mair 1984	45%	×
Chernykh, Doyle, and Power 2017	28.7%	✓
Coma and van Ham 2015	29.5%	✓
Dahlberg et al. 2013	41.1%	✓
David and Holliday 2012	71%	×
Gervasoni 2010	81%	×
Huber and Inglehart 1995	40%	×
Kato and Laver 1998	28%	×
Kato and Laver 2003	17%	×
Kerby and Blidook 2014	28%	✓
Laver 1998	63%	×
Lupu and Michelitch 2018	46%	✓
McElroy and Benoit 2007	67%	✓
McLean et al. 2009	33.6%	✓
O'Malley 2007	60%	×
Pétry, Collette, and Klingemann 2012	21%	✓
Polk et al. 2017	39.7-42.9%	×
Ray 1999	45%	×
Ray and Narud 2000	72%	×
Rohrschneider and Whitefield 2007	42%	✓
Schmitt and Loughran 2017	30%	✓
Szöcsik and Zuber 2015	32.21%	×
Warwick 2005	23.1%	×

C Figures

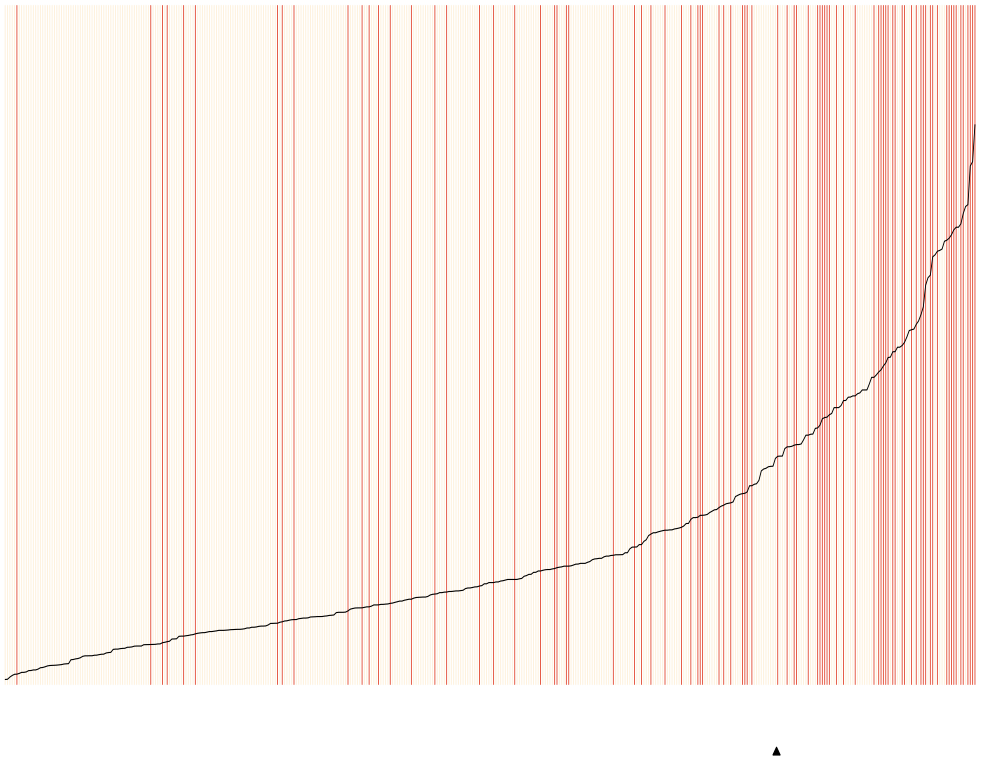

Figure 3: Separation Plot (based on Model 1)