**BRINGING THE PUBLIC BACK IN: SURVEY EXPERIMENTS AND FOREIGN POLICY ANALYSIS**

**Jonny Hall**

**London School of Economics and Political Science**

**j.hall5@lse.ac.uk**

*Apologies for the lack of proper references and a bibliography! This quickly written paper is the product of quite a lot of reading (80 articles that use a survey experiment), but to be honest the contributions have become less clear to me over time. I think the overall thrust of my argument is probably pretty obvious (though note this is very much not my standard area of research), but I’d really appreciate any feedback on whether anything works well/less well in terms of the more forward-looking content of the essay. Do I need more on what FPA is, for example?*

**INTRODUCTION**

Several studies have noted the significant growth of ‘population-based survey experiments’ in political science [Mutz 2011], and especially with regards to public opinion on foreign policy issues [Dafoe et al. 339; Brutger et al. 2022; Kertzer 2023]. Between 2020 and 2024, the *Journal of Conflict Resolution* alone published 27 articles including a survey experiment of the U.S. public, amounting to just over 10 percent of all published articles in that period.[[1]](#footnote-2) These articles have engaged with already well-established research agendas in the experimental study of U.S. public opinion on foreign affairs, including support for the use of nuclear weapons [Rathbun and Stein, 2020; Allison et al. 2022; Press et al. 2013; Dill et al. 2022], the moral foundations that influence foreign policy stances [Powers et al. 2022; Kertzer and Zeitzoff 2017; Kertzer et al. 2020; Brutger and Rathbun 2021], and the presence of audience costs [Casler and Clark 2021; Tomz 2007; Levendusky and Horowitz 2012; Levy et al. 2015]. With increasingly feasible access to representative samples of relevant populations and the ability to manipulate specific variables, these survey experiments mean that we know more than ever about how the U.S. public views the world and the underlying logic of these opinions. Whilst these pieces often acknowledge concerns regarding the external validity of these surveys [i.e. whether these findings extrapolate beyond their experimental setting; Morton and Williams 2010, 255], this article puts forward two interrelated arguments related to how the findings of this research agenda translate to the ‘real world’. Firstly, with a focus on the experimental survey works on audience costs,[[2]](#footnote-3) the article posits that experimental design choices result in a lack of ecological validity [i.e. the similarities between the research environment and the context the experiment is meant to be mimicking; Morton and Williams 2010, 265]. A lack of ecological validity is not in itself problematic but leads here to overstated effect sizes due to unrealistic informational environments and a shortfall of relevant priors. Secondly then, this article argues that the relationship between public opinion and foreign policy is under-theorised, particularly in terms of how elites *understand* and *utilise* ‘public opinion’. As such, the article closes with a call for a closer relationship between survey experiments and the International Relations (IR) subfield of Foreign Policy Analysis (FPA), arguing that FPA both offers models for considering exactly *how* public opinion influences foreign policy and would also be improved using survey experiments.

**SURVEY EXPERIMENTS AND VALIDITY**

Internal validity refers to the ‘*approximate truth*’ of the findings within the (in this case) experimental context [Morton and Williams 2010, 255, italics in original]. Put another way, it relates to the claims of the experimental results in that setting. This contrasts with external validity which, as mentioned above, refers to the validity of the knowledge claim outside the experimental context. External validity is explored in two ways in this article: ‘whether a causal relationship holds (1) over variations in persons, settings, treatments that *were* in the experiment and (2) for persons, settings, treatments, and outcomes that *were not* in the experiment’ [Shadish et al. 2002, 83]. In that sense, external validity cannot strictly be judged in terms of the research itself, but rather in terms of the claims surrounding an experiment’s findings [Brewer and Crano 2014, 11]. One could side with Mook’s [1983] ‘defense of external invalidity’ for experiments: test results should be assessed on whether they support the specified theory or not, rather than on account of the generalisability of their findings.[[3]](#footnote-4) For example, Stanley Milgram’s infamous obedience experiment at Yale University in 1961 was not meant to be an accurate representation of the typical contexts where subordinates felt compelled to follow orders [Mook 1983, 385]. On the other hand, plentiful research has shown the importance of context in influencing experimental results, such as how participants adapt their behaviour simply because they are being monitored [known as the ‘Hawthorne effect’; McDermott 2011, 37; Mutz 2011, 13/4. Either way, most experiments ultimately wish to explain real life phenomena (as Mook 1983, 386 acknowledges). Even if ‘most experiments are highly localized and particularistic’, experimental researchers nearly always want to generalise [Shadish et al. 2002, 19], and that is no different for the research studied here.

How, then, can external validity be attained? Whilst external validity is inevitably going to include human judgments as to the generalisability of the claims being made, the burden of *proof* for external validity is always empirical support over multiple studies [Shadish et al. 2002, 34, 96; Morton and Williams 2010, 266]. Accordingly, it is a mistake to think of external validity as the property of one experiment [Shadish et al. 2002, 96]. Because validity relates to inferences and not methodology, it is also problematic to follow conventional wisdom and assume that there is an inherent trade-off between internal and external validity [Shadish et al. 2002, 34, 97], such as Tomz’ [2007, 825] reference in his landmark study of audience costs to how ‘experiments are vulnerable on the dimension where observational data are most compelling: external validity’. Much like a well-executed field experiment, internally valid experimental survey findings from a well-specified population sample could very logically translate to other similar populations [Mutz 2011, 131].

A final prominent concern relates to ecological validity, which is seen as an ‘aspect’ or ‘form’ of external validity [Kihlstrom 2021, 468; Brewer and Crano 2014, 19]. However, ecological validity does not refer to the generalisability of the results per se, instead describing the similarities between the research environment and the environment in which experimental findings are theorised to apply [Morton and Williams 2010, 265].[[4]](#footnote-5) After all, an experimental setting with high ecological validity for one particular context could reduce external validity by producing results specific to a certain context [Morton and Williams 2010, 265]. As Verba [1964, 502 in Renshon 2015, 670] strongly put it, ‘the experimental model does not need to ‘look like’ the real world’. Yet, as is argued below, there is a relationship between ecological and external validity, particularly concerning the generalisability of large effect sizes that may not emerge in the real world. Accordingly, although the survey experiment literature studied here normally merges ecological and external validity [Renshon 2015, 667], this article recognises the differences between the two terms to fully flesh out the critiques related to the real world application of the research agenda under study.

**SURVEY EXPERIMENTS AND OVERSTATED EFFECTS**

With a particular (but not exclusive) focus on the survey experiment literature on audience costs, this section argues that the effects of experimental treatments on foreign policy stances are oftentimes exaggerated because of external and ecological validity issues related to survey experiments. Firstly, survey respondents are often asked to consider a policy issue ‘as if’ they were policymakers [for example, Allison et al. 2022, 1776]. As Mutz [2021, 232] has claimed, this is itself an ‘experimental manipulation’ that can change how information is received and which policies are supported. Meta-analysis has shown how respondents increasingly turn to abstract (or ‘high-level construal’) thinking when instructed to place themselves in someone else’s shoes [Soderberg et al. 2015], consequently exaggerating their own biases [Renshon 2015, 661]. Construal level theory also shows the discounting of future costs by survey participants, resulting in higher levels of support for hypothetical uses of military force than when asked to simply *consider* casualty numbers before responding [Huddleston 2019].

Secondly, there are broader issues related to the ecological validity of the predominant treatment in this literature: vignettes. In a study of the comparative accuracy of five different experimental treatments [Hainmueller et al. 2015], the single vignette design fares worst. On the one hand, Mutz [2021, 231] argues that these largely neutral treatments fall short in their manipulation effects of the independent variable, thus decreasing their experimental realism (the extent to which the desired experimental effects mirror the target context). For example, Huff and Kertzer’s [2018, 62] vignette for their conjoint survey experiment on how the public defines terrorism is quite brief:

The incident: a shooting.

The shooting occurred at a church in a foreign democracy with a history of human rights violations.

There were two individuals killed in the shooting.

The shooting was carried out by a Muslim individual with a history of mental illness.

News reports suggest the individual had been in an ongoing personal dispute with one of the targets.

Brief experimental treatments are particularly prone to under-recognised problems of informational equivalence, such as how labelling a country a ‘democracy’ can affect respondents’ beliefs about the geography of said country, hence bringing any manipulation effects into question [Dafoe et al. 2018, 399].[[5]](#footnote-6) On the other hand, adding too much contextual detail can reduce treatment effects because of the inability of respondents to recall the relevant information [Brutger et al. 2022]. Consider the experimental treatment in Dill et al.’s [2022, 14] study of support for the use of nuclear weapons, which are news article articles between 650 and 700 words long ‘intended to heighten the realism of the experience for subjects, increasing the external validity of any findings’:



This statement makes the problematic claim concerning the external validity of one experiment and appears to conflate external validity with ecological validity [Carpenter and Montgomery 2020, 83], but even on that final point there are grounds for debate, particularly in terms of providing survey respondents with specific figures for ‘probability of destroying target’, ‘estimated Libyan civilian deaths’, and the casualty figures for American civilians if the terrorist attack was carried out.

Indeed, the broader issue here, which is particularly pertinent in the context of foreign policy issues, is that survey experiments ‘typically obliterate the distinction between the supply of information … and its consumption’ [Kinder 2007, 157]. Although the Almond-Lippmann consensus on public opinion (i.e. that it is disinterested, fickle, and irrelevant) has long been debunked by the ‘rational public’ perspective [Holsti 1996; Page and Shapiro 1992; Hurwitz and Peffley 1987], is it also true that levels of knowledge of foreign affairs in the U.S. remain low and oftentimes are not factually accurate [Boettcher and Cobb 2006, 842; Eichenberg 2015; Guisinger 2017; Allison et al. 2022, 1784]. For example, in 2015 42 percent of polled Americans answered that the U.S. had discovered weapons of mass destruction in Iraq [Carpenter and Montgomery 2020, 148], whilst in 2020 and 2021 25 percent of the U.S. public were not aware that the country was still involved in the Afghanistan War [Kriner and Kreps 2021]. There are well-detailed shortfalls of this kind of ‘quiz show’ questioning [Russett 1990, 89 in Nincic 1992, 778], but for democratic responsiveness to occur in mechanisms such as audience costs, the public must hold knowledge of foreign policy issues and corresponding elite positions [Guisinger 2009, 535; Kertzer et al. 2023, 1]. As such, the fact that respondents are directly provided information on comparatively specific issues ranging from the effectiveness of weapon types [Press et al. 2013; Sagan and Valentino 2017; Dill et al. 2022; Rosendorf et al. 2023] and successful drone strikes [Boddery and Klein 2021, 1] to covert operations against other democracies [Carnegie et al. 2023] and politicians’ stances on security issues [Tomz et al. 2020] is potentially problematic in terms of ecological validity. Iyengar’s [2000] study that the effects of political advertisements were made redundant by allowing participants a television remote that enabled them to simply change channels is a telling example of the significance of uninterrupted information supply to a captive audience when considering treatment effects [McDonald 2020, 270]. Furthermore, one-off experimental treatments struggle to capture the long-term and social nature of political stimuli [Gaines et al. 2007, 5/6] which is important here given Kertzer and Zeitzoff’s [2017, 547-9] work on the importance of peer cues in opinion formation on foreign policy.[[6]](#footnote-7) Direct information is not in itself a problem, but low levels of ecological validity can create challenges for significant claims about external validity.

Also unlike the messy nature of the ‘real world’, experiments oftentimes have no relevant priors that ground attitudes in everyday politics, thus leading to overstated effect sizes [Barbaras and Jerit 2010, 239; McDonald 2020, 268; Croco et al. 2020, 9; Croco et al. 2022, 365/6]. Put another way, when more background knowledge is available, attitudes are harder to change. McDonald [2020, 278/9] shows via the use of ‘mirror experiments’ that there is a 14-point reduction in the audience costs suffered between a hypothetical Democrat and Barack Obama when replicating Tomz’s [2007] landmark study. As McDonald [2020, 279] argues, ‘[b]y making the experiment more ecologically valid, we now fail to find an effect that for a decade has been largely accepted in the literature on international conflict.’ There are also issues with the measurement of audience costs, as scholars have argued that answers related to a *situational* approval question (‘[d]o you approve, disapprove, or neither approve nor disapprove of the way the US president handled the situation?’) can represent a drop in *general* approval which are necessary for audience costs to function (i.e. ‘a drop in approval ratings has real and significant consequences for leaders’ reelection prospects’; Levendusky and Horowitz 2012, 328 in Croco et al. 2020, 13).[[7]](#footnote-8) This lack of priors and conflation of situational and general approval has led experimental research to predict unlikely drops in general approval for presidents that back down in an international crisis, such as 22 and 24 percent [Levy et al. 2015 995; Trager and Vavreck 2011, 534]. Similarly to the above, repeat experiments that focus on general approval with real politicians find significantly reduced effects given that foreign policy is just one component of presidential approval [Croco et al. 2020, 10].

There is also the well-documented problem of the relationship between self-stated attitudes and observable political behaviour, such as the nearly 20 percent difference between reported and actual turnout in the 2012 and 2016 American National Election Studies [Peterson et al., 242]. The challenges of observing real world political behaviour are particularly challenging in this area, given the complexity of the relationship between experimental treatments on foreign policy issues and choices made by individuals at the ballot box [Lupton and Webb 339/40]. Though proponents of audience costs theory (and experiments more generally) point to the strategic selection of policymakers as evidence of why we are unlikely to see audience costs emerge [Tomz 2007, 825], the problems outlined above in terms of ecological and external validity may contribute to the shortage of concrete examples of this much-discussed phenomena. The empirical record suggests that leaders typically put forward ambiguous threats in international crises, rather than necessarily boxing themselves in [Snyder and Borghard 2011, 439]. Indeed, experimental research has shown that vague statements, explanations of policy changes, and the invocation of relevant foreign policy themes can all reduce (and in some cases completely nullify) audience costs [Trager and Vavreck 2011; Levendusky and Horowitz 2012; Queck and Johnston 2017]. This aligns with other survey experiment works on foreign policy issues that point to the importance of elite rhetoric and framing in influencing policy stances, such as support for lethal drones and autonomous weapons systems [Kreps 2014; Kreps and Wallace 2016, 840; Rowling and Blaukwamp 2021, 1; Rosendorf et al. 2023, 3]. Beyond rhetoric, experimental research has also argued that certain types of war financing give leaders greater domestic leeway when pursuing foreign policy strategies [Flores-Macias and Kreps 2017, 977].

In sum, these different factors attest to how ‘‘[b]acking down’ in an international crisis is what leaders and their domestic publics *make out of it*’ [Quek 2022, 1968, emphasis added]. Yet, there is very little conceptualisation of the exact causal mechanism of audience costs [Slantchev 2006, 445/6; for a similar critique of bottom-up theories of international trade policy, see Guisinger 2009]. This is troublesome because of the variability and interactivity of international crises and the accompanying domestic politics. Indeed, Tomz [2007, 825] recognised that ‘the best way to make progress on complicated topics is to analyze data from multiple sources’ and that ‘the experiments in this article provide new insights to *complement* what others have found with historical data’ [emphasis added]. Likewise, Trager and Vavreck’s [2011, 19] early adaptation of Tomz’s work referred to their experiments ‘as *complements* to, rather than substitutes for, historical analysis’ [emphasis added]. ‘[A]ny healthy research program ought to avoid relying too much on one particular method’ [Kertzer et al. 2023] because overreliance on a single method restricts the scope of questions that are asked and number of explanations that are explored [Kinder 2011, 527; Mearsheimer and Walt 2013, 448; Hyde 2015, 411; Lupton and Webb 2023, 350], but this is arguably what has happened in the contemporary ‘cottage industry’ of audience costs studies [Baum and Potter 2014, in Kertzer and Brutger 2016, 246].[[8]](#footnote-9) That is, works begin with the premise that audience costs do exist and make the findings narrower with each new survey experiment, rather than necessarily linking back to real world data. Levendusky and Horowitz [2012, 325], for example, refer to how Tomz’s study showed ‘that the public *does actually impose audience costs* on democratic leaders’, whilst Trager and Vavreck [2011, 535] claim that Tomz demonstrated that ‘audience costs *exist*’ [emphases added]. Given that the effect of individual (and sometimes contrasting) variables is likely exaggerated due to the stylised one-off experimental treatment of audience costs literature [Trager and Vavreck 2011, 533/4; Evers et al. 2019, 443], Walt and Mearsheimer’s [2013] broader lament concerning hypothesis testing is relevant here: ‘[i]f several published articles on a given topic all contain statistically significant but substantively different results … how do we decide which one to believe?’ The following section details how the survey experiment literature on foreign affairs has dealt with the generalisability of their experimental findings.

**SURVEY EXPERIMENTS AND THE RELATIONSHIP BETWEEN PUBLIC OPINION AND FOREIGN POLICY**

The survey experiment literature concerning public opinion on foreign affairs takes various approaches to delineating the relationship between public opinion and foreign policy. Some works spend little to no time spelling out the impact of public opinion on foreign policy behaviour, either in terms of overall impact or the scope conditions where we would expect to see the public’s views influence decision-making [inter alia, Kertzer et al. 2014; Levy et al. 2015; Kertzer and Brutger 2015; Kertzer 2020; Kertzer et al. 2020; Brutger and Rathbun 2021]. Although these studies understandably avoid the problem of overstating the influence of public opinion on foreign policy outcomes, Kleinberg [2023, 373] has argued that ‘the “so what” question … looms large’ when the relevance of public opinion to policymaking is not specified. Consequently, various pieces by Tomz and Weeks [2013, 850; 2017, 850; 2020, 182; 2021, 1] include clear and detailed engagement with IR and FPA works on both the public opinion-foreign policy relationship alongside the relevance of public preferences in that specific foreign policy area. These references cover the multitude of ways in which public opinion can affect policymakers’ preferences, ranging from direct electoral pressures to the need for governments to overcome institutional hurdles and mobilise resources. Given theories about the influence of the ‘attentive public’ on (foreign) policy outcomes [Almond 1960; Rosenau 1961], the authors also usefully specify the views of the most politically active segments of the population sample [Tomz 2007, 832; Tomz and Weeks 2013, 861; 2020, 189]. Several pieces on international military crises spell out how public opinion has previously influenced policymakers in that area [Rathbun and Stein 2020, 789-790, 810-811; Kertzer et al. 2021, 315; Lee 2022, 983], whilst the strongest pieces outline the specific mechanisms that they see public opinion influencing foreign policy [Grieco et al. 2011, 577-580; Barnhart et al. 2020, 636]. Tomz et al. [2020, 120], for example, outline two pathways that public opinion could influence foreign policy: a) direct responsiveness of policymakers to public opinion and b) support for politicians with similar foreign policy attitudes at the ballot box.

 Other works make similar claims about the relationship between public opinion and foreign policy, but oftentimes these complex and contingent dynamics are oversimplified [as they often have been by IR theories despite the ‘domestic turn’; Kaarbo 2015, 198-199]. This is particularly noticeable in the survey experiment literature on public opinion regarding the use of nuclear weapons. Press, Sagan, and Valentino [2013, 193/4] fairly recognise that although FPA literature disagrees on the extent to which public opinion influences decision-making, scholarship has increasingly rejected the adage ‘that politics ends at the water’s edge’, hence making ‘public opinion a reasonable place to test theories of nuclear non-use’. In a similar vein, Sagan and Valentino [2017, 41] claim that their findings ‘provide insights into whether the U.S. public would be a constraint or a goad to encouraging a president to use nuclear weapons’. These statements clearly speak to the underlying logic of theories such as audience costs when it comes to democratic foreign policymaking, but increased specificity would point towards how accountability environments vary by policy area [Staniland and Narang 2018]. As Kreps and Wallace [2016, 832] contend, it is not necessarily obvious why policymakers would need prospective support for one-off airstrikes as in these experiments.[[9]](#footnote-10) Indeed, the authors effectively recognise this when they note that their retrospective experiment (i.e. respondents react to news of a nuclear strike they have been informed has already happened) ‘is probably closer to the way the public would learn about an attack in the real world’ [Press et al. 2013, 197/8]. Given that Tannenwald’s ‘nuclear taboo’ thesis invokes a shared social norm [Press et al. 2013, 193], surveying the American public on their views of nuclear weapon usage makes perfect sense as an exercise in theory testing. However, to advance this research agenda, the ecological validity of these experiments must be increased, such as exposing respondents to the inevitable military and political objections that would occur in a publicly discussed nuclear strike [Carpenter and Montgomery 2020, 165/6].[[10]](#footnote-11) Additionally, the relevance of public opinion on nuclear weapons to policymakers must be specified further beyond statements such as ‘we focus on democratic states because mass public opinion is more likely to influence national policy in such states’ [Dill et al. 2022, 13].

 Related to this discussion of nuclear weapons is the relevance of the populations studied by most of the literature studied here, given that foreign policy decisions are made by elites and not the general public. This is not to say that the public should not be participants in survey experiments, given that they offer the opportunity to interrogate the microfoundations of theories in IR [Kertzer 2017] such as tripwire effects [Musgrave and Ward 2023; Allison et al. 2022], the effect of international organisations on support for the use of force [Grieco et al. 2011; Tomz and Weeks 2021], and the ‘democratic advantage’ in crisis bargaining [Kertzer et al. 2021]. There are also articles where public *perceptions* (rather than the impact of these perceptions) are the phenomenon being studied, such as survey experiments on identifying terrorism [Huff and Kertzer 2018] and attributing cyber attacks [Musgrave and Leal 2022].[[11]](#footnote-12) In all these cases, the ‘population’ of ‘population-based survey experiments’ seems to be directly relevant to the phenomenon being investigated [Mutz 2011]. However, this does not take away from the fundamental insight of FPA: that foreign policy interactions are grounded in human decision-makers, thus meaning that the study of foreign policy should be actor-oriented and actor-specific [Hudson 2005, 1, 4]. Contrary to the idea that elites are ‘generally busy, wary of clinical poking, and skittish about revealing information about their decision-making processes and particular choices’ [Hafner-Burton et al. 2013, 368 in Dietrich et al. 2021, 598], literature reviews have demonstrated the feasibility of elite samples, with the mean response rate of targeted at 40 percent [Dietrich et al. 2021; Kertzer and Renshon 2022, 534].

 FPA research has shown that elites and the public have different stances on foreign policy issues because of variety in experience [Busby et al. 120], though Kertzer’s [2022; see also Renshon et al. 2023] recent meta-review has documented how the psychological and ideological differences between elites and publics have been overstated in the conventional wisdom. If members of the public and elites behave similarly, this significantly advances the external validity of studies such as Kertzer et al.’s [2022] study of groupthink dynamics in the international arena. The problem, however, is that survey and experimental research also show that elites consistently misperceive the content of public opinion [Kertzer et al. 2023]. Because the literature above pointing towards the link between public opinion and foreign policy behaviour implicitly relies on elites accurately *knowing* the content of public opinion, this elite-public perception gap has sizeable consequences ‘for everything from our theoretical models of the domestic politics of conflict, to the empirical strategies we use to test microfoundations in IR’ [Kertzer et al. 2023]. Rather than treating the public as a source of information for analysing ‘leaders’ political incentives’ [Tomz and Weeks 2013, 850], the key variable becomes ‘elite perceptions of public opinion’ [Kertzer et al. 2023]. This important insight should encourage scholars to pay more attention to the nuances in the relationship between public opinion and policymakers, especially considering how some previous experimental studies are couched in the paradoxical ideas that a) public opinion *does influence* decision-makers and b) policymakers misperceive public opinion [Press et al. 2013, Gelpi et al. 2009].

This oversight from the survey experiment literature is particularly striking given that the perception gap between elites and actual public opinion is well documented in the U.S. context [Kell and Destler 1999; Page and Bouton 2006, 221; Gift and Monten 2021]. Concerning foreign affairs, FPA literature points to nuances in perceptions and usage of ‘public opinion’ by policymakers. Studies argue that U.S. elites believe that public opinion should be considered in foreign policy decision-making [Powlick 1991, 625], such as the use of opinion polls to guide policy [Hinckley 1992]. However, the ‘public opinion/foreign policy nexus’ is rarely so linear or unidimensional [Powlick and Katz 1998]. Considering opinion polls for example, elites rarely initiate polls on foreign policy issues [Kull and Destler 1999, 11, 208], and when such data is available, policymakers can use these as strategic tools to dictate issue salience or guide persuasion attempts for a public that is largely deemed to be unwise and disinterested in foreign policy [Powlick 1991, 630; Powlick 1995, 439; Powlick and Katz 1998, 30; Kull and Destler 1999, 208; Druckman and Jacobs 2015]. As an alternative to polls, policymakers have significant leeway in the sources of public opinion that they can turn to [Powlick 1991, 630]. Foyle’s [1999, 21] study of U.S. presidents found personal perceptions of public opinion to be significantly more influential in foreign policy decision-making than specific opinion polls, whilst Pew Research Center data on the sources of public opinion used by congresspeople, presidential appointees, and senior civil servants likewise pointed to the insignificance of opinion polls in decision-making [Kull and Destler 1999, 212]. Instead, policymakers have been found to largely rely on Congress and the media for perceptions of public opinion on foreign policy issues [Powlick 1995; Kull and Destler 1999, 208].[[12]](#footnote-13) Given how policymakers closely engage with congresspeople and the media [Kull and Destler 1999, 211], alongside the importance of these actors in making foreign policy issues electorally salient [Powlick and Katz 1998], there is an understandable logic for policymakers to use these proxies [Powlick 1995, 435-7]. As such, it is not that foreign policymakers do not consider public opinion, but rather that perceptions of ‘public opinion’ are complex, especially given that leaders are oftentimes focused on ‘latent opinion’ that does not yet concretely exist [Zaller 1992]. Due to the generally low electoral salience of foreign policy issues, there is also little incentive for policymakers to gain an accurate reading of public opinion [Kull and Destler 1999, 230-1]. This complexity and contingency is little discussed in the survey experiment literature, but as noted above, is very significant for the assumptions of this research agenda.

The importance of public opinion for foreign policy decisions is ultimately an empirical question, which fortunately the survey experiment literature has explored. Hainmueller et al.’s [2015] innovative study on the naturalisation of immigrants in Switzerland referenda shows promising degrees of congruence between experimental treatments and voter behaviour,[[13]](#footnote-14) but it is not clear how this translates to citizen behaviour concerning more distant foreign policy issues. Tomz et al.’s [2020] survey experiments pointed towards the importance of security issues in influencing voter choice in the U.S. and Israel, though other studies have found contrasting results in the U.S. [Guisinger 2009], Pakistan [Clary and Siddiqui 2021], Japan [Horiuchi et al. 2018], and the European Union [Oktay 2018]. Like the above, this is a varied and contingent relationship that survey experiments on foreign affairs cannot be expected to account for completely, but Aldrich et al. [1989, 132] remains a relevant guide for thinking about the conditions in which foreign affairs influence voter choice: voters must have coherent foreign policy attitudes that are salient in an election and relate to different candidates. It is not clear that many of the survey experiments cited here would necessarily meet these criteria. The next two sections provide some suggestions for resolving the critiques of the survey experiment literature posited here.

**MOVING FORWARD: EXPERIMENTAL DESIGN AND ENGAGEMENT WITH FOREIGN POLICY ANALYSIS**

This section points to best practices in survey experiment work that advance the external and ecological validity of this research. The employment of multiple surveys enhances the external validity of the results by demonstrating the consistency of findings across different contexts. Resource constraints are clearly relevant here, but Renshon et al.’s [2023] six survey experiments across four different countries on the credibility of democracies’ resolve represent the par excellence of this research agenda. Another excellent use of multiple experiments in political science comes from Robison’s [2022, 557-560] study of countervailing political messages, as the author uses three experiments varied in degrees of abstraction to demonstrate the external validity of the claims being made despite the changing degree of priors held by respondents [see also Croco et al. 2022]. As Robison study shows, the ability of experimental surveys to quickly incorporate contemporary events represents a great strength of the method [Mutz 2011], and especially in terms of external and ecological validity, such as reducing informational equivalence problems [Dafoe et al 2018, 413]. Experimental research using real world examples have demonstrated reduced political costs for foreign policy inconsistency between policies and rhetoric in comparison to previous experimental studies [McDonald et al. 2019; Croco et al. 2020]. As some of the most prominent exponents of survey experiments have recently argued, ‘there are significant benefits to designs that leverage real world events and actors and allow us to calibrate the magnitude of effects found in highly abstract experiments where observers have no real priors about the context’ [Kertzer et al. 2023].

If incorporating contemporary events is not possible, the external validity of experiments is most likely only going to be improved by attempts to make experimental treatments as realistic as possible. To give but a few examples, Krebs and Wallace [2016, 835] explicitly design their vignettes to mirror debates over drone strikes, Musgrave and Ward [2023, 10] vary troop levels to match the variation of U.S. overseas military contingents, and Huff and Schub [2018] base the speeches used in their survey experiment on examples from previous U.S. presidents. There are also options to move beyond the predominant experimental treatment of vignettes given some of the pitfalls raised above in terms of ecological validity. Political science survey experiments use different media formats such as photographs and videos [Mutz 2011, 62/3; Mutz 2021, 221], as well as allowing respondents to choose the media content they are exposed to [Peterson et al., 244]. Given the essential (and changing) role that the media plays in how citizens are exposed to foreign policy issues [Baum and Potter 2019], increased consideration of the media in survey experiments would enhance the research’s ecological validity. Looking beyond variation in treatments, behavioural games used in the experimental survey literature on foreign affairs offer an excellent example of alternative experimental methods that likely enhance external validity [Hafner-Burton et al. 2014; Renshon 2015; Kertzer et al. 2022]. In a similar logic, studies that allow for open-ended responses may allow for increased understanding of respondents true ‘opinions’ on foreign policy issues [Dill et al. 2022; Dill et al. 2023; Carnegie et al. 2023; Powers and Renshon 2023, 739]. This is because

[p]eople do not make simple statements, they shade, modulate, deny, retract.… These manifestations of uncertainty are just as meaningful and interesting as the definitive statements of a belief system. [Hochschild 1981, 238 in Carpenter et al. 2021, 920]

Here, then, researchers are better able to ‘considerations’ that could be generally relevant across different foreign policy contexts rather than ‘definitive attitudes’ which may be less accurate [Zaller and Feldman 1992, 585 in Carpenter et al. 2021, 920].

In terms of population samples, I have argued above that due to the ultimate nature of foreign policy decision-making, elites are an extremely relevant population for survey experiments on this topic. In the U.S., survey experiments have explored ‘elite’ views on cooperation with international law [Hafner-Burton et al. 2014], the impact of status on decision-making [Renshon 2015], the influence of multilateral approval on support for the use of force [Busby et al. 2020], perceptions of public opinion on foreign policy issues [Kertzer et al. 2023], and how this influences support for the use of force [ibid; Lin-Greenberg 2021]. There are some debates about the ‘eliteness’ of these samples [Kertzer and Renshon 2022, 538], but the same cannot be said for an impressive dataset from the Israeli Knesset that has been used to explore the impact of personal traits on the perceived credibility of international actions [Yarhi-Milo et al. 2018], the perceived resolve of democracies [Renshon et al. 2023], and the impact of voter preferences on elites’ willingness to support the use of force [Tomz et al. 2020]. This research provides important examples of the feasibility of elite survey experiments but also should provide a reminder that elite samples do not necessarily resolve generalisability issues [Kertzer and Renshon 2022, 531], given that the findings of the Knesset sample may not translate beyond that context [Dietrich et al. 2021]. As such, one can question Dill and Schubiger’s [2021, 612/3] claim that ‘public opinion is crucial for war-related decision-making in the United States’ is supported by references to Tomz et al. [2020; in that their article focuses on Israel] and Tomz and Weeks [2013; in that the piece looks at public opinion *on* foreign policy, rather than its impact].

 Whether elite samples are possible or not, the survey experiment literature on foreign affairs would be improved by increased use of examples and targeted evidence to demonstrate the real world effects of survey-based phenomena. Whilst survey experiments are often cited as a means to circumnavigate endogeneity problems surrounding unpopular decisions, it is not necessarily the case that policymakers do not articulate this line of thinking. Through careful archival work, Payne [2023, 1-2] provides primary evidence of multiple U.S. presidents explicitly admitting to decisions that put presidential popularity above military calculations, such as Richard Nixon and Henry Kissinger’s plan in 1972 to ‘find some formula that holds [Vietnam] … together a year or two’ even if the overall U.S. strategy was doomed. Indeed, case study analysis enables Payne [44-5] to convincingly argue that conceptualising electoral pressures as a ‘contingent pressure’ helps explain the contradictory findings of previous scholarship that either stresses or downplays foreign policy decision-making constraints in democracies. By contrast, Carnegie et al. [2023, 258] claim that their survey results on public views of covert operations suggest that leaders ‘may select covert missions against democracies because the public prefers them for instrumental reasons (despite also valuing transparency)’, as opposed to the extant literature which points towards using secrecy to avoid public criticism. This is fundamentally an empirical question that cannot be expected to be answered in detail in a journal article including a detailed survey experiment, but surely exploring empirical material related to the examples of covert operations listed in the piece would be a worthy endeavour as a type of plausibility probe [Carnegie et al. 2023, 236]? Considering Hyde’s [2015, 412] argument that the value of survey experiments becomes particularly clear when integrated into manuscript-length projects that allow for more thorough engagement with theory and empirics to enhance external validity, the comparative shortfall of books that significantly rely on survey experiments on U.S. foreign policy issues is striking [see Kertzer 2016; Renshon 2017; Walsh and Schulzke 2018; Rathbun 2023]. Clearly, it is not my place to insist that exponents of experimental surveys write manuscript-length books, but I believe it will be a worthy endeavour to help increase the coherence and scholarly impact of this research agenda.

 In terms of engagement with relevant theory, FPA scholarship offers multiple different avenues for development. This aligns with Kaarbo’s [2015, 189] conceptualisation of a ‘psychologically--oriented and agent-based’ FPA as a ‘crucible’ for advancing the ‘domestic turn’ in IR, alongside continuing the growing conversation between psychological approaches and foreign policy studies [Kertzer 2023]. At the most fundamental level, FPA provides a reminder of the *contingency* and *interactivity* of the impact of public opinion on decision-makers which, as argued above, is oftentimes overlooked in the survey experiment literature.

Firstly, survey experiments could be used to explore individual-level variation among leaders and the potential impact this has on foreign policy outcomes. In the U.S. context, Foyle [1999, ix-x] has documented how the influence of public opinion is moderated by variation between presidents on a) the perceived value of incorporating public opinion into decision-making and b) the decision context in which choices are made. As Foyle [1999, 281] noted, the different views of U.S. presidents on the importance of public opinion means that audience costs are leader-contingent. Survey experiments could be utilised to explore this elite-level variation, adding nuance and depth to Foyle’s work by exploring *when* and *how* leaders consider or discard public opinion. It is one thing to say that public opinion influences policymaking, but another to think about whether this acts as a constraint, a guide, or a means of generating support [see Foyle 1999, 262]. As Kertzer et al. [2023] note, exploring the relationship between belief systems and perceptions of public opinion would be a worthy endeavour to better comprehend the elite-public perception gap discussed previously. Understanding elite-level variation concerning the influence of public opinion would speak to FPA theories on these issues like poliheuristic theory [Mintz 2004; Vijayalakshmi 2017] and two-level games [Putnam 1988; Conceicao-Helt and Mello 2017]. This research would also align with FPA’s emphasis on leader-specificity, which links to the potential of using survey experiments to explore personality traits as an alternative to the at-a-distance methodology of Leadership Trait Analysis [Hermann 1980]. FPA research has shown that specific leadership traits influence both foreign policy decision-making processes and outcomes [inter alia, Dyson 2006, Holmes and Yarhi-Milo 2017], and these traits could be identified or tested for their effects experimentally. Given that previous research has demonstrated that politicians consistently have different personality traits to the public [Dal Bó et al. 2017 in Kertzer and Renshon 2022, 536], these discrepancies could be explored at a mass and elite level.

Secondly, the survey experiment literature could engage with FPA works on domestic-level influences of foreign policy. At the broadest level, the interaction between regime type and the influence of public opinion on foreign policy could be explored further. Although this article has deliberately focused on the research relating to U.S. public opinion on foreign affairs, this does at least partially reflect the U.S.-centrism of this research agenda. Given that research suggests that non-democracies are more sensitive to public opinion than the traditional democracy/autocracy dichotomy suggested, studies should follow in the footsteps of those that have already explored public opinion on foreign affairs in non-democratic, hybrid, and transitional regimes [Kertzer 2023]. If practical concerns do not transpire, discerning citizens’ views on the influence of public opinion on foreign policy in both nondemocracies and democracies could be a fruitful research avenue on the ‘public opinion/foreign policy nexus’.

Short of that, FPA offers multiple perspectives for conceptualising this nexus and the influencing factors of note. Thomas Risse-Kappen [1991] pointed to domestic political structures and coalition-building processes as influencing factors for the impact of foreign policy, whilst Baum and Potter argue [2015] that multiparty systems lead to increased democratic accountability than two-party systems. These differences are noteworthy for claims such as Dill et al.’s [2023] argument that different attitudes towards force protection are reflected in American, British, and Israeli military doctrines, given that public opinion could well be exercising different degrees of influence across these states. As Staniland and Narang [2018] argue in the Indian context, there are varying ‘accountability environments’ for different foreign policy issues depending on the ‘clarity of responsibility’ and ‘salience’ of the policy. This points to how the varying importance of public opinion across different policy issues should be more fully considered by the survey experiment literature but also suggests room for surveying elites on whether they perceive public opinion to be more or less relevant for specific policy areas. Powlick [1991, 622; 1995, 434], for example, found that officials working on more specialised foreign policy issues referred to the challenges of even gauging what ‘public opinion’ would constitute in their issue area. Generally, the survey experiment literature overlooks the importance of salience in the ‘public opinion/foreign policy nexus’, such as how the influence of public opinion varies based on the stages of the election cycle and decision-making process [Payne 2023; Knecht and Weatherford 2006]. Baum and Potter’s [2008] ‘elasticity of reality’ framework that charts the changes in elite and public information levels concerning military conflicts offers an excellent framework for survey experiments to better incorporate the variation in the influence of public opinion over time.

Beyond the direct impact of public opinion, FPA research on other influences of foreign policy could work in coordination with survey experiments. Studying the U.S. Congress, Gelpi and Grieco [2015] show how public disapproval of foreign policy actions has ‘competency costs’ that result in increased challenges to a president’s domestic legislative agenda. Accordingly, not only could survey experiments explore congressional views on foreign policy and their impact on decision-makers [Saunders 2023], but also how foreign policy failures or success will likely influence vote choice on domestic issues. Following Kreps’ [2010] study into the impact of elite consensus on divergent public opinion on troop commitments in Germany, Canada, France, and Italy, survey experiments could investigate whether elites believe themselves to be insulated from the consequences of unpopular decisions when prominent parties agree on foreign policy issues. Another set of actors that could be surveyed relates to the ‘sub-bureaucratic politics perspective’ developed in the U.S. foreign policy context [Gvosdev et al. 2019]. This approach argues that FPA has not paid enough attention to the ‘lower bureaucratic echelons’ and how this area influences the creation of foreign policy decisions, which provides a clear link to political science research that has studied how institutions shape policymaking procedures and policy options [Grose 2021, 153-4]. Along the lines of non-experimental survey research that studies the varying preferences of foreign aid officials between countries [Dietrich 2018], variance within and between different state bureaucracies could be explored to contribute to the bureaucratic politics paradigm in FPA [e.g. Kaarbo 1998; Gulen 2022].

Lastly, experiments could be used in conjunction with role theory from FPA. Cantir and Kaarbo [2012, 5] called attention to how a state’s role ‘may be contested both vertically (between elites and masses) and horizontally (among elites)’, which could be tested at both levels via survey experiments. Open-ended questions would offer an opportunity for the significance of role-based thinking at both an elite and mass level. In addition, survey experiments (or behavioural games) across different countries would offer the opportunity to explore the interactivity of ‘role-taking’ and ‘role-making’, as per McCourt’s [2012] ‘interactionist’ approach to role theory. At the broadest level, survey experiments offer opportunities for increased conversations between constructivism and FPA [for this argument, see Houghton 2007], given the ability of the method to point to the causal significance of different frames of foreign policy issues [for example, Boettcher and Cobb 2006; Kreps and Wallace 2016; Guisinger and Saunders 2017; Huff and Kertzer 2018]. It would also align with FPA’s long-standing engagement with psychological literature on information processing that was founded on experimental methods [Hudson 2005, 10-12]. The literature on historical analogies [Khong 1992] and populism [Destradi, Cadier, Plagemann 2021] represent potentially fruitful research avenues for further exploration of the significance of social images in foreign affairs via survey experiments.

**CONCLUSION**

As shown by the number of works cited in this piece alone, the survey experiment literature concerning public opinion on foreign affairs represents a vibrant part of contemporary IR. If Holsti was once able to refer back to the ‘Almond-Lippmann consensus’ on public opinion on foreign policy, it now seems like the opposite is true: the public’s views are deemed a highly relevant topic of scholarly inquiry. Utilising technological developments and the strengths of experimental methodology, this research agenda has advanced our understanding of the influencing factors of public opinion on foreign policy issues, particularly in terms of the importance of core values. To be clear, this is a worthy endeavour: public opinion is an important factor in the formulation and implementation of foreign policy [Holsti 2006], and survey experiments offer a powerful tool for ‘inferring how public opinion works in the real world’ [Gaines et al. 2007, 4].

 However, this article has also argued that this literature does not represent a panacea to understanding the complex relationship between public opinion and foreign policy in the ‘real world’. Firstly, the article argued although survey experiments need not mirror the real world (and this could theoretically decrease external validity in some cases), there is a need to meaningfully consider the ecological validity of survey experiments more fully, especially concerning the consequences of this for effect sizes. McDonald’s [2020] reassessment of the audience costs survey experiment literature with an increased focus on ecological validity points to the importance of this consideration. Secondly, the article argued that the survey experiment literature must pay more attention to the specificity and variation in the relationship between public opinion and foreign policy. As Knecht’s [2010, 3] model of ‘conditional responsiveness’ aptly summarises in the U.S. context, ‘presidents sometimes lead, sometimes follow, and sometimes ignore the American public.’ FPA’s emphasis on contextuality and contingency highlights the dangers of making external validity claims based on single experiments, as the audience costs literature has sometimes done. This problem is particularly striking in the context of the elite-public perception gap recently detailed by Kertzer et al [2023]; if elites do not actually know public opinion, then it is unclear how exactly mass opinion influences policy. To that end, this article has called for more survey experiments with foreign policy elites for an improved understanding of the most relevant population in foreign policy decision-making. As Hyde [2015, 410/1] put it, ‘the relevance of public opinion should be predicated on specific hypotheses or empirical implications rather than on generalizations by issue area.’ In addition, survey experiments on foreign affairs should increasingly engage with real world scenarios to enhance external validity, both in terms of vignettes and theory validation. Beyond the experiments themselves, this article has called for an increased conversation between the survey experiment literature and FPA research on elite variation, issue salience, other actors in the foreign policymaking process, and role theory. In that way, FPA can both *enhance* the survey experiment literature’s external validity and *benefit from* the method by improving the empirical support for its concepts and theories.

1. This article focuses on the U.S., in part because this reflects the focus of most of the extant work, but also because of feasibility issues (in both the survey experiment literature and FPA), having read 70 articles that utilise survey experiments on the U.S. public. As FPA scholars have recognised [see, for example, Blummer 2021], this is problematic for generalising beyond the American context. [↑](#footnote-ref-2)
2. The article focuses upon this literature as it is referred to as the gold standard of this research agenda and speaks to the broader logic underlying much of this work: that citizens have (in)direct influence over their state’s foreign policies. [↑](#footnote-ref-3)
3. For similar defences in a foreign policy context, see Horowitz and Levendusky [2012, 327/8] McLean and Roblyer [2017, 242/3]. [↑](#footnote-ref-4)
4. For a broader definition of ecological validity as ‘whether an effect has been demonstrated to occur under conditions that are typical for the population at large’, see Brewer and Crano (2014, 21). [↑](#footnote-ref-5)
5. Only 14 percent of the articles studied in Dafoe et al. [2018, 399] recognise the potential pitfalls of informational equivalence. [↑](#footnote-ref-6)
6. Even here, however, Kertzer and Zeitzoff’s [547-549] argument shows that peer cues matter more when it comes to scenarios including ‘the deployment of special forces to combat terrorists in the Middle East’. Given the generally covert nature of special operations forces and the low salience of U.S. counterterrorism campaigns [Hall 2021], the ecological validity of these findings can also be questioned. [↑](#footnote-ref-7)
7. Asking respondents for their opinion on how a leader handled an issue also increases the likelihood of a participant voicing a critical perspective [Croco et al. 2020, 9]. [↑](#footnote-ref-8)
8. Noticeably, Carpenter et al. [2021, 912] also use the ‘cottage industry’ label regarding the literature on citizens’ attitudes towards international law and war policies. [↑](#footnote-ref-9)
9. Whilst I agree with their critique, I find Kreps and Wallace’s [2016] argument that prospective approval would be more relevant for drone strikes unconvincing [Hall 2022]. [↑](#footnote-ref-10)
10. There is an irony here in that Sagan and Valentino [2017, 45/6, 61, 76] justify this position by aiming to replicate the same trade-offs that Harry Truman and the U.S. public faced when considering the use of nuclear weapons at the end of World War II, thus overlooking the sizeable normative changes (even if not meeting the standard of a ‘taboo’). In response to Carpenter and Montgomery’s critique, Sagan and Valentino [2020, 177/8] later ran a survey experiment in which the Chairman of the Joint Chiefs of Staff flagged the (il)legality of a proposed nuclear strike, but even here there are informational equivalence and ecological validity issues given that the Joint Chiefs of Staff are proposing an illegal option, which could suggest the gravity of the situation [Carpenter and Montgomery 2020a, 184; Bell 2023, 169]. [↑](#footnote-ref-11)
11. Outside of the U.S. context, the survey experiment literature on the effects of counterinsurgency operations has been a particularly fruitful research avenue [see Kertzer 2023]. [↑](#footnote-ref-12)
12. Bernard Cohen [1973]’s study pointed to the significance of networks of friends, family, and academics for perceptions of public opinion, whilst Kull and Destler [1998, 209] also point to the importance of the “vocal public”. [↑](#footnote-ref-13)
13. The experiments fare less well when it comes to predicting the absolute levels of rejection rates in the real referendums, with a 16 percent difference. [↑](#footnote-ref-14)