Rethinking the Tradeoffs of Within-Subjects Designs in International Relations Experiments

Clara H. Suong[*] Scott Desposato[†] Erik Gartzke[‡]

July 2, 2024

Abstract

All experimentalists in International Relations face a choice over their designs—including that between between-subjects and within-subjects designs. This manuscript helps them make informed decisions in two ways. First, we reassess existing findings on the advantages of within-subjects designs, such as higher statistical power and precision in treatment effect estimates. Second, we contextualize the tradeoffs of within-subjects designs by examining their external validity across countries. Our results are based on a unique survey experiment with within-subjects and between-subjects designs in Brazil, China, Japan, and Sweden and a meta-analysis of 26 paired within-subjects conditions from existing lab experiments. We find that within-subjects designs provide higher statistical power, precision in treatment effect estimates, insights on treatment effect heterogeneity, and strong external validity across countries. However, we find evidence of order effects in within-subjects experiments. Our results imply that within-subjects designs are a practical option in International Relations—particularly for under-resourced researchers—albeit with tradeoffs.

Word count: 10,239
Keywords: experiment; experimental design; within-subjects design; survey experiment; lab experiment
Competing interests: The author(s) declare none.
Data availability statement: The data used for the analyses in this paper will be made publicly available upon publication.

[*] Virginia Tech. e-mail: clara.suong@vt.edu, web: https://www.clarahsuong.com/
[†] University of California, San Diego. e-mail: swd@ucsd.edu, web: http://swd.ucsd.edu
[‡] University of California, San Diego. e-mail: egartzke@ucsd.edu.

Table of Contents

1 Introduction

Many experiments in IR (International Relations) feature one type of experimental design—the so-called "post-only," between-subjects design. Alternative designs, such as the within-subject design, have been much less common (Lupton and Webb, 2022) but are being increasingly adopted by IR researchers (e.g. Kertzer, Renshon and Yarhi-Milo, 2021; Myrick, 2020, 2023; Renshon, Yarhi-Milo and Kertzer, 2022; Tomz and Weeks, 2013; Tingley and Walter, 2011).

In this research note, we reassess and contextualize existing scholarship on experimental designs, focusing its findings on the tradeoffs of within-subjects designs (Chaudoin, Gaines and Livny, 2021; Clifford, Sheagley and Piston, 2021; McDonald and Hanmer, 2023; Transue, Lee and Aldrich, 2009) We extend this scholarship by examining the pros and cons of using within-subjects designs in IR experiments, including both survey and lab experiments. Specifically, we highlight the pros and cons of using within-subjects designs in IR experimental research in two ways: with a unique survey experiment in Brazil, China, Japan, and Sweden that feature a within-subjects design and a between-subjects design, both of which vary the same experimental factor, thereby allowing us to compare the two designs by their measurements of the same effect; and with a meta-analysis of selected within-subjects designs in lab experiments. Using these data on survey and lab experiments, we examine the magnitude, precision, heterogeneity, order effects, and external validity of treatment effects captured in within-subjects designs.

We find that within-subjects designs outperform between-subjects designs in statistical power, precision, and the ability to examine heterogeneous treatment effects, consistent with existing findings (Clifford, Sheagley and Piston, 2021). Our results also suggest that within-subjects designs do not fall behind between-subjects designs in cross-country external validity for the direction of the treatment effects. However, our results suggest the existence of order effects in within-subjects designs, similar to the studies by Chaudoin, Gaines and Livny (2021) and McDonald and Hanmer (2023) and consistent with Transue, Lee and Aldrich (2009)'s evidence on the order effects in multiexperiment surveys. Our results suggest that within-subjects designs are a practical alternative to between-subjects designs in IR experiments, their statistical power rendering them particularly useful to under-resourced researchers, but also are vulnerable to order effects.

In the following section, we first discuss the choice of survey designs faced by experimental IR scholars, focusing on the choice between between-subjects designs and within-subjects designs. We then show the "lay of the land," examining the prevalence of within-subjects designs among existing works of experimental IR. We also classify popular within-subjects designs into three types and discuss each type's attributes. This is followed by a discussion of the possible advantages and disadvantages of using them in IR research. Next, we provide an overview of our experiment and meta-analysis and discuss the results of our analyses. We then conclude.

2 Between-Subjects Designs and Within-Subjects Designs

Experimentalists studying individuals' opinions and behavior in IR face choices over various designs in their experiments (Mintz, Yang and McDermott, 2011). One of the decisions they face is over how to randomly assign subjects to experimental conditions and when and how to measure their responses to the conditions. Often this comes down to choosing to assign subjects to the treatment and control conditions via a between-subjects design, a within-subjects design, or a combination of the two (Dietrich, Hardt and Swedlund, 2021; McDermott, 2002; Kertzer and Renshon, 2022).

Most experimental work in IR employs between-subjects designs (Lupton and Webb, 2022)—especially the "post-only" in which the outcome is measured only once. These designs are notable for the single exposure of subjects to experimental conditions and the single measurement of the outcome of the conditions (Clifford, Sheagley and Piston, 2021). In other words, they first randomly and simultaneously assign subjects to one of the conditions—the treatment and control conditions—and then measure their responses to the conditions once for each of the subjects.

Repeated-measures designs—in which the outcome is measured more than once—can be a good alternative to between-subjects designs (McDonald and Hanmer, 2023). These include within-subjects designs in which "each subject is exposed to all experimental conditions and the dependent variable is measured after each condition" (Clifford, Sheagley and Piston, 2021, 1050) and which are alternatively referred to as cross-over or cross-subject designs (e.g. Kertzer and Zeitzoff, 2017; Tingley, 2014).[1] In other words, within-subjects designs differ from between-subjects designs in that subjects in the former are assigned to multiple conditions per experiment—as opposed to one condition in the latter—and given all conditions of an experimental factor/variable. In such designs, each subject's exposures to all conditions are implemented sequentially and often in a randomized order.[2] Thus, subjects usually do not differ from each other from (the number of) the conditions they are assigned to but in the order in which they are exposed to the conditions.

While between-subjects designs are much more common in IR, scholars have noted the (potential) usefulness of the less popular within-subjects designs. For example, Tingley (2014) notes the "great promise in the class of cross-over designs, which involve treatment assignment in multiple stages" (449) and urges scholars to adopt them more frequently in their research.

3 Within-Subjects Designs in IR Research

When are within-subjects designs used in IR research? Although not as popular as between-subjects designs, within-subjects designs have not been rare among IR experiments. Substantively, they have

[1] Other repeated-measures designs include (quasi) pre- and post-designs (Clifford, Sheagley and Piston, 2021) and between-subjects designs repeatedly applied to a panel of respondents (e.g. Gartner, 2008; Huff and Schub, 2018; Kertzer, 2016; Tingley, 2011). Conjoint designs are also reliant on within-participant comparisons (Clifford, Sheagley and Piston, 2021; Mutz, Druckman and Green, 2021) to which our logic is applicable. However, we focus on standard, non-conjoint within-subjects designs instead of conjoint designs as an alternative to between-subjects designs for two reasons: the former is much easier to understand and to implement than the latter; the key estimand for the former and between-subjects designs is the average treatment effect (ATE) whereas the latter's key estimand is the average marginal component effect (AMCE), not ATE, making the comparison between conjoint and between-subjects designs difficult.

[2] Some within-subjects designs do not randomize the order of the conditions because they include control conditions as the baseline and initial condition to which the following treatment conditions are compared to (e.g. Bush and Prather, 2021; Demarest, Jost and Schub, 2024; Naoi, Shi and Zhu, 2022; Spilker, Nguyen and Bernauer, 2020; Yarhi-Milo, Kertzer and Renshon, 2018; Yarhi-Milo and Ribar, 2022).

been particularly popular among the two strands of research often reliant on small samples: works on individuals' strategic interactions that use lab experiments (e.g. Adamson and Kimbrough, 2022; Chaudoin, Hummel and Park, 2024; Chaudoin and Woon, 2018; Fahoum, Pick and Shamay-Tsoory, 2023; Erev and Rapoport, 1990; Hundley, 2019; Moxnes and van der Heijden, 2003; Slusher, Rose and Roering, 1978; Tingley and Walter, 2011; Wilson, 1969); and studies on opinions and beliefs of elites that use survey experiments (e.g. Demarest, Jost and Schub, 2024; Hafner-Burton, LeVeck and Victor, 2015, 2017; Kertzer, Renshon and Yarhi-Milo, 2021; Kertzer and Zeitzoff, 2017; Naoi, Shi and Zhu, 2022; Rathbun, Kertzer and Paradis, 2017; Renshon, Yarhi-Milo and Kertzer, 2022; Yarhi-Milo, Kertzer and Renshon, 2018).[3]

Methodologically, within-subjects designs in IR can be classified into three types: single-module, multiple-module, and repeated-wave designs. Letting a module M consist of the minimum number of the waves (also often referred to as rounds/periods/repetitions/plays/iterations/games) during which a subject is sequentially exposed to all conditions for a within-subjects experimental factor T, we can describe the three categories as follows:

1. Single-module, within-subjects designs: Within-subjects designs in which all subjects are exposed to all conditions for T throughout the experiment and to each condition of T only once;

2. Multiple-module, within-subjects designs: Within-subjects designs in which all subjects are exposed to all conditions for T throughout the experiment and to multiple modules for T;

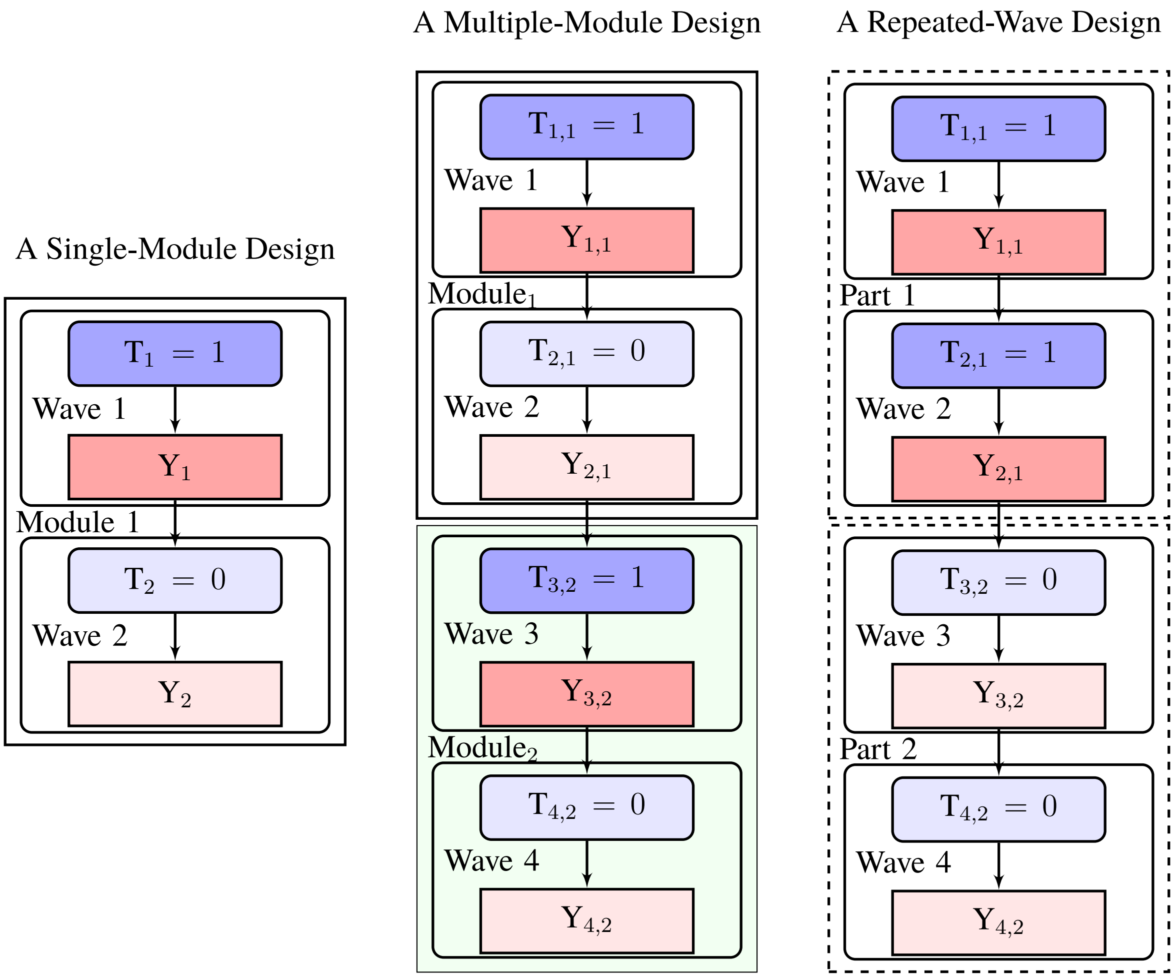
3. Repeated-wave, within-subjects designs: Within-subjects designs in which all subjects are exposed to each condition of T successively over repeated waves

Figure 1 illustrates the difference between the three types of within-subjects experimental designs, displaying the order of items an example subject would receive in example single-module,

[3] Conjoint experiments also rely on within-subjects comparisons (e.g. Avey et al., 2022; Kertzer, Renshon and Yarhi-Milo, 2021; Huff and Kertzer, 2018; Lim and Tanaka, 2022; Majnemer and Meibauer, 2023) but are not a focus of this research note.

multiple-module, and repeated-wave within-subjects designs with two levels/conditions for one within-subjects factor.

Figure 1: Example Within-Subjects Designs



In the example, T_w for the single-module design ($\mathrm{T}_{w,m}$ for the multiple-module design and $\mathrm{T}_{w,p}$ for the repeated-wave design) represents the condition assigned to a respondent in Wave *w* (in Module m or in Part p) for an experimental factor/variable. T_w, $\mathrm{T}_{w,m}$, and $\mathrm{T}_{w,p}$ can be 0 or 1. $\mathrm{Y}_{i,w}$ ($\mathrm{Y}_{w,m}$ or $\mathrm{Y}_{w,p}$) denotes the outcome variable that records the response of the subject to condition T_w (or $\mathrm{T}_{w,s}$ or $\mathrm{T}_{w,p}$) in Wave *w* (in Module m or in Part p).

All three designs are subject to a within-subjects constraint on assigning a respondent to conditions : $\mathrm{T}_1 = 1 - \mathrm{T}_2$ (or $\mathrm{T}_{1,m} = 1 - \mathrm{T}_{2,m}$ or $\mathrm{T}_{1,p} = \mathrm{T}_{2,p} = 1 - T_{3,p+1} = 1 - T_{4,p+1}$). In other words, this requires that the condition assigned to the respondent in Wave 1 to differ from the condition she is assigned to in Wave 2 for the single-module design (within each Module *m* for the multiple-module design) or that the conditions assigned to the respondent are the same only within Part p for the repeated-wave design.

The three designs differ in terms of the number of times and the sequence in which each respondent is exposed to the conditions. In the single-module design, she is exposed to both experimental conditions–1 and 0–only once. In the multiple-module design, she is exposed to a module of two conditions before being exposed to another module of the two conditions again. In the repeated-wave design, she is exposed to both conditions but each of the conditions is immediately repeated before the exposure to the other condition. In other words, she is first exposed to condition 1, followed by another wave of condition 1 before being exposed to condition 0 twice.

Table 1 displays a list of recently published articles in IR by their type of within-subjects designs and design features. We focus on the IR articles recently published in the following journals: *American Political Science Review*, *American Journal of Political Science*, *Journal of Politics*, *International Organization*, *International Studies Quarterly*, *Journal of Conflict Resolution*, *Journal of Peace Research*, or *Journal of Experimental Political Science*.

Table 1: Types of Within-Subjects Designs in Recent Experimental IR Research

Type	Number of Within-Subject Factors	Order of the Key Within-Subjects Conditions Randomized / Counterbalanced	Studies
Single-Module	One	Randomized	Myrick (2020, 2023); Kertzer and Zeitzoff (2017); Renshon, Yarhi-Milo and Kertzer (2022); Tomz and Weeks (2013); Touchton et al. (2020)
		No	Böhm, Fleiß and Rybnicek (2021); Bush and Prather (2019, 2021); Crandall et al. (2016), Chaudoin, Hummel and Park (2024),[5] Demarest, Jost and Schub (2024), Fahoum, Pick and Shamay-Tsoory (2023)[6], Hafner-Burton, LeVeck and Victor (2015, 2017); Kertzer, Renshon and Yarhi-Milo (2021); Naoi, Shi and Zhu (2022); Son and Park (2022); Spilker, Nguyen and Bernauer (2020); Tingley and Walter (2011); Yarhi-Milo, Kertzer and Renshon (2018); Yarhi-Milo and Ribar (2022)
Multiple-Module	Multiple	Randomized Only for One Factor	Böhm, Fleiß and Rybnicek (2021)
		No	Chaudoin and Woon (2018)[7]
Repeated-Wave	One	Counterbalanced	Adamson and Kimbrough (2022), Hundley (2019); Moxnes and van der Heijden (2003)
		No	Chaudoin and Woon (2018)*

Several interesting observations can be made here. First, most within-subjects designs are single-module, exposing a respondent to all conditions for a factor only once. Second, most studies do not (or cannot) randomize or counterbalance the order of conditions.[4] Third, multiple-module designs often involve multiple within-subjects factors that are manipulated whereas the other two types involve one within-subjects factor (sometimes combined with other between-subjects factors).

4 Assessing Within-Subjects Designs

4.1 Within-Subjects Designs' Advantages

While between-subjects designs are simple to design and understand, they have a big downside: relatively low statistical power and low precision of the treatment effect estimates (Mutz, 2011). Consequentially, researchers are pressed to increase statistical power by recruiting a bigger sample of respondents (Mutz, 2011) or by including control variables (Bowers, 2011), such as respondents' demographic information, in the design and analysis. However, the former requires resources and the latter can be ineffective at increasing precision due to low correlations between dependent and control variables (Montgomery, Nyhan and Torres, 2018; Mutz, 2011; Sheagley and Clifford, 2023).[5]

In turn, within-subjects designs have several advantages over between-subjects designs. They have higher statistical power and higher precision for the effect estimates (due to the correlation

[4] This will be discussed in depth in the following section on order effects.

[5] Only Experiment 1 of Fahoum, Pick and Shamay-Tsoory (2023) is a (mixed) within-subjects design; Experiment 2 is purely a between-subjects design.

[6] Only Study 2 of Crandall et al. (2016) qualifies as a within-subjects design.

[7] Chaudoin and Woon (2018)'s 4 "sessions" with the combination of BN-BF and CN-CF treatments in Part 1 and Part 2 are a multiple-module, within-subject design with 2 factors varied within-subject independently—the presence of feedback (or lack thereof) and the high/low valuations. The high-low valuation conditions are assigned within-subject in repeated "rounds" across Part 1 and Part 2. However, the feedback conditions (BF and CF treatments) are assigned subjects in Part 2—which always occurs after Part 1—and no feedback conditions in Part 1 during these 4 "sessions." In the remaining 6 "sessions", in contrast, the same high-low valuation conditions are repeated by part (in Part 1 and Part 2) for each individual. However, the order of conditions does not differ by individual, due to the lack of counterbalancing.

[5] C.f. See Klar, Leeper and Robison (2020) for a different perspective.

between the two dependent variables) and, consequentially, require smaller samples (Clifford, Sheagley and Piston, 2021; Kane, 2024; Kertzer and Renshon, 2022; McDonald and Hanmer, 2023; Mutz, Druckman and Green, 2021). They also enable researchers to examine the heterogeneity in treatment effects, offering insights into the variation in treatment effects (Clifford, Sheagley and Piston, 2021).

4.2 Within-Subjects Designs' Disadvantages

However, two concerns may prevent IR researchers from using within-subjects designs.

Order and Learning Effects First, IR researchers may be concerned about the order effects in within-subjects designs (e.g. Chaudoin and Woon, 2018; Hundley, 2019; Kertzer and Zeitzoff, 2017; Naoi, Shi and Zhu, 2022; Renshon, Yarhi-Milo and Kertzer, 2022; Son and Park, 2022; Tingley, 2011; Tingley and Walter, 2011). Within-subjects designs expose subjects to multiple conditions sequentially, unlike between-subjects designs which expose them to one randomly chosen condition. Thus, this brings the possibility of the order of conditions administered to the subjects affecting the estimates of average treatment effects (Chaudoin, Gaines and Livny, 2021; Transue, Lee and Aldrich, 2009).[6]

The order effects in within-subjects designs that may render estimates of treatment effects with lower accuracy stem from two factors: consistency pressures; and demand (or learning) effects. Consistency pressures refer to the onus on individuals to appear consistent by sticking to their initial answers given during the first condition and not changing their answers during subsequent conditions. Because individuals link consistency to a positive self-image (Cialdini, Trost and Newsom, 1995), asking the same questions multiple times in these designs may induce the subjects to stay committed to their initial answers and "anchor" them (McDonald and Hanmer, 2023) even if

[6] Order effects can also arise from the order of the questions—whether they appear before or after treatments—biasing treatment effect estimations (Blackwell et al., 2023; Clifford, Sheagley and Piston, 2021; Klar, Leeper and Robison, 2020; Montgomery, Nyhan and Torres, 2018; Sheagley and Clifford, 2023).

the conditions they are assigned to change. Thus, consistency pressures in within-subjects designs may induce the subjects to be insensitive to treatments and result in smaller effect sizes (McDonald and Hanmer, 2023).

In contrast, demand or learning effects, if existing, may induce subjects in within-subjects designs to be overly sensitive to treatments (Mummolo and Peterson, 2019).[7] Works on survey experiments note that measuring the same dependent variables multiple times in within-subjects designs may signal to subjects the design and goals of the experiment, and lead them to "learn" about them and be sensitive to the researchers' intentions and behave accordingly (Clifford, Sheagley and Piston, 2021; McDonald and Hanmer, 2023; Tingley, 2011). Existing studies on strategic interactions using lab experiments and student samples also note the possibility of subjects learning (or strategically adjusting) via the order and repetition of multiple experimental conditions (e.g. Cheung and Friedman, 1997, 1998; Muller et al., 2008; Ostrom, Walker and Gardner, 1992; Slonim and Roth, 1998; Tingley, 2011). The strategic games played in lab experiments are structured so that those who are sensitive to experimental conditions are financially rewarded. This incentive structure of lab experiments can result in a large demand or learning effects particularly because they often involve student samples. The initially "naive" or "novice" students in the lab experiments (Kertzer, 2022, 542) may learn to be strategic by repeatedly playing similar games over time, resulting in increasing treatment effects over repetitions.

In both survey and lab experiments, order effects may shift the treatment effects and hinder an accurate inference about the effect size. Both consistency pressures and demand effects have the potential to hinder an accurate inference about the effect size in within-subjects designs but in opposite directions (McDonald and Hanmer, 2023). Consistency pressures, if they exist, may lead to smaller treatment effects for within-subjects designs than those for between-subjects designs. In contrast, a demand or learning effect would result in larger treatment effects for within-subjects designs. If neither exists, treatment effects for within-subjects designs would not differ from those

[7] C.f. Mummolo and Peterson (2019) find no evidence of demand effects in their between-subjects experiment.

for between-subjects designs.

When employing within-subjects designs, IR scholars can use several tactics to mitigate the order effects: to combine them with between-subjects assignments of other factors (e.g. Bush and Prather, 2021; Crandall et al., 2016; Kertzer, Renshon and Yarhi-Milo, 2021; Renshon, Yarhi-Milo and Kertzer, 2022; Yarhi-Milo, Kertzer and Renshon, 2018); to complement a within-subjects experiment with a separate between-subjects experiment for the same factor (e.g. (Naoi, Shi and Zhu, 2022; Son and Park, 2022)); or to randomize or counterbalance the order of within-subjects treatments (e.g. Adamson and Kimbrough, 2022; Hundley, 2019; Moxnes and van der Heijden, 2003; Myrick, 2020, 2023; Renshon, Yarhi-Milo and Kertzer, 2022).[8]

Counterbalancing refers to adopting an experimental design that assigns subjects to difference sequences of experimental conditions. This includes the "complete a full A-B/B-A experimental design" (Tingley and Walter, 2011, 21) in which a randomly chosen half of the subjects are exposed to experimental condition A and then condition B and the other half to condition B and then condition A for a factor with levels A and B. For example, when Hundley (2019) exposed all subjects to 2 experimental conditions (playing five-round games and 10-round games), she used "counterbalancing to prevent order effects," ascertaining that "[f]or every subject who played the five-round games before the 10-round games, there was a different subject who played the 10-round games first." In other words, the order of the key within-subjects experimental factor—playing long or short games—was implemented as an independent between-subject factor; all subjects played both five-round and 10-round games but differed in the order of the long and short games they played.

However, there remain several issues with these tactics. First, an experiment combining within-subjects assignments of a factor with between-subjects assignments of other factors may be still vulnerable to the order effects from the former. Additionally, running 2 separate experiments with within-subjects and between-subjects manipulations requires additional resources.

[8] The third option is recommended by Chaudoin, Gaines and Livny (2021).

Furthermore, randomizing or countervailing the order of conditions, while recommended by existing scholarship (Chaudoin, Gaines and Livny, 2021), is not feasible for all studies. This is particularly true for the studies with within-subjects designs for which one of the conditions is the "baseline" or control condition (e.g. Bush and Prather, 2021; Demarest, Jost and Schub, 2024; Naoi, Shi and Zhu, 2022; Spilker, Nguyen and Bernauer, 2020; Yarhi-Milo and Ribar, 2022).[9] Because the controlled conditions establish the baseline, these studies assign subjects to the controlled conditions first and then other treatment conditions afterwards. In particular, priming studies are subject to this constraint because it is difficult to "undo" priming; once a researcher introduces subjects to a stimulus, asking them to not think about it afterwards is very difficult.[10] This implies that priming studies' estimates of treatment effects may include the design-induced order effects from having the controlled condition first and may differ from estimates from between-subjects designs. Moreover, even if the order of conditions is randomized and orthogonal to other variables, the order effect may still emerge in within-subjects designs, which may affect treatment effects (Chaudoin, Gaines and Livny, 2021), and differentiate their estimates from between-subjects' estimates of the same effects.

Cross-National Generalizability In addition to possible order effects, IR researchers may hesitate from adopting within-subjects designs due to the possible design effects' on the generalizability of the experimental results across different countries. IR experiments often involve fielding across multiple countries (Bassan-Nygate et al., 2023) because testing IR theories can involve a comparison of multiple countries or an examination of relations between them (e.g. Tomz and Weeks 2013). Thus, an IR researcher may be understandably concerned about within-subjects designs' potential to affect the generalizability of treatment effects, the effects' directions in particular (Bassan-Nygate et al., 2023), across countries.

[9] Many of these use single-module designs in Table 1.

[10] Framing studies in contrast are more receptive to randomizing and counterbalancing. "Undoing" is less of an issue in framing studies, in which subjects are introduced to different stimuli and expected to compare them (Naoi, Shi and Zhu, 2022, 6).

5 Research Design and Data

Noting these potential advantages and disadvantages of choosing within-subjects designs over between-subjects designs in IR research, we employ two strategies to compare the two. First, we field our own experiment that compares the two designs. Second, we complement this with a meta-analysis of existing IR studies with within-subjects designs. Each of these strategies is discussed in detail below.

5.1 Our Experiment

Our experiment incorporates both between-subjects and within-subjects designs, with both designs manipulating the same factor. We then compare the precision, accuracy, direction, and generalizability of average treatment effects of the same experimental factor measured in the between-subjects and within-subjects designs of our own survey experiment. In particular, we compare the most common type of within-subjects designs in IR research—the single-module within-subjects design—with a between-subjects design.

Our survey instrument includes a vignette and questions on subjects' support for the use of force and demographic and attitudinal formation. The vignette features a crisis involving a hypothetical neighboring country's pursuit of nuclear weapons and respondents' support for the use of force in the crisis. In the survey, respondents are first instructed to read a short scenario ("vignette") about potential conflicts between two hypothetical countries ("Country A" and "Country B").[11] The primary treatment featured in the vignette is whether or not the United Nations has authorized Country A to use force against Country B. Our substantive hypothesis is that IO approval will (positively) affect individual support for the use of force, based on existing literature.[12]

Our subjects are exposed to 2 waves (a single module) of a within-subjects factor's conditions

[11] Respondents thinking of specific cases in hypothetical scenarios per se do not negate the existence of treatment effects in survey experiments (Brutger et al., 2023; Suong, Desposato and Gartzke, 2023).

[12] An additional treatment was a dichotomous assignment of the regime type of the target country, Country B, as democratic or not democratic. We focus on this treatment in another paper.

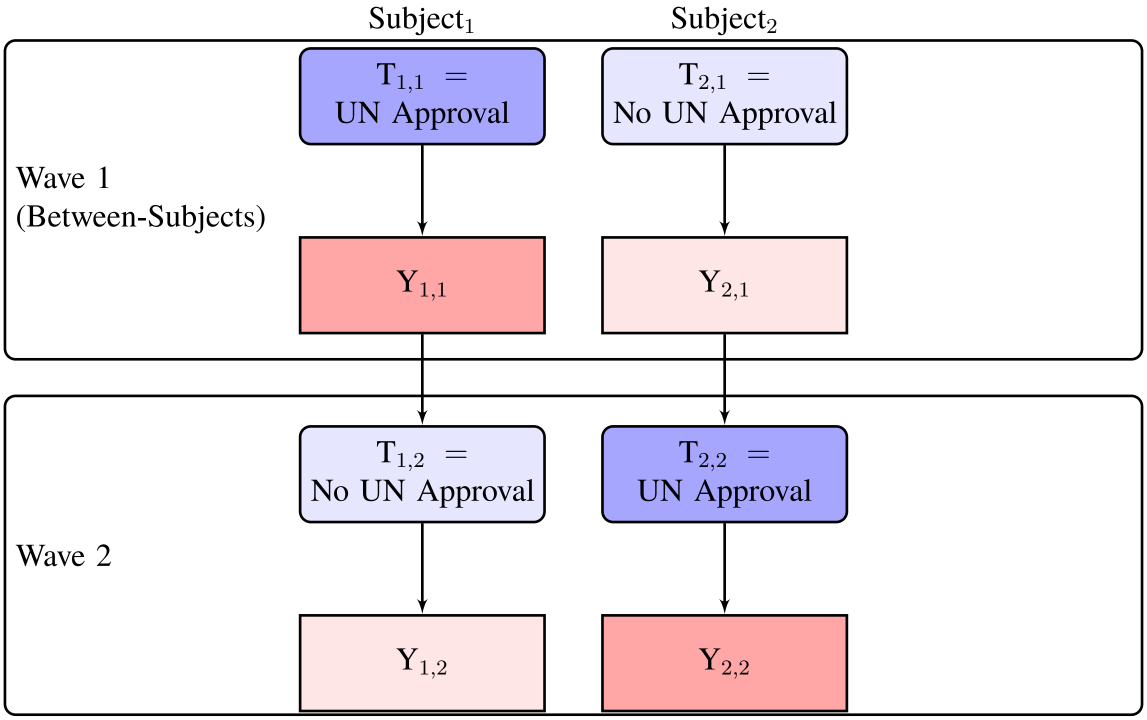
on UN approval. Following the vignette, subjects are asked “Should Country A attack and use force to resolve the situation?” Subjects can respond to this key question with either “attack” or “not attack.” These items constitute the treatment-outcome pair (Wave 1). After answering this question, they receive a follow-up question, which varies the presence or lack of UN approval in the opposite direction (Wave 2). Specifically, subjects who have seen a vignette where the UN did (or did not) authorize the use of force in Wave 1 are asked whether they would support the use of force if the UN were not (or were) to grant authorization in Wave 2. In other words, each respondent is exposed to both treatments of UN approval and of its absence over Waves 1 and 2 but the order in which she receives them is randomized.

Effectively, this design provides within-subjects treatments where the approval of the UN is varied via the vignette in Wave 1 and via the follow-up question in Wave 2. At the same time, Wave 1—in which the subjects are first assigned to a condition—is a post-only, between-subjects design. Thus, our design is a within-subject design with an embedded between-subjects design in Wave 1.

Figure 2 demonstrates our experiments with within- and between-subjects designs for example respondents, Subject_1 and Subject_2, in two waves—Wave 1 and Wave 2. Each wave *W* consists of an experimental condition for the factor UN Approval Subject *i* is assigned to ($\text{T}_{i,w}$) and an outcome measure ($Y_{i,w}$).[13]

[13] See Appendix D.3 for the wording of the vignette and questions.

Figure 2: Our Within-Subjects Design for Two Example Subjects, Subject$_1$ and Subject$_2$



Note: $T_{i,w}$ represents the conditions randomly assigned to respondent *i* in Wave *w* for the factor/variable UN approval. $T_{i,w}$ can be 0 (No UN Approval) or 1 (UN Approval). $Y_{i,w}$ denotes the outcome variable that records whether respondent *i* supports the use of force in Wave *w*. $Y_{i,w}$ can take the value of 0 (Oppose War) or 1 (Support War). Note that our design imposes two constraints on assigning respondent *i* to conditions: $T_{i,1} = 1 - T_{i,2}$; and $P(T_{i,1} = 1) = 0.5$. The within-subjects design imposes the constraint $T_{i,1} = 1 - T_{i,2}$, meaning that the condition assigned to respondent *i* in Wave 1 differs from the condition *i* is assigned to in Wave 2. The between-subjects design imposes the second constraint that the probability that respondent *i* is assigned to the UN approval condition (or the No UN Approval condition) in the first wave is 0.5 for all *i*.

Figure 2 illustrates the within-subjects element of our experiment. In Wave 1, Subject$_1$ is assigned to the UN Approval condition and her response to UN approval is recorded. In her Wave 2, she is assigned to the other condition/level for the variable UN approval—the No UN Approval condition—after which her response to the No UN Approval condition is measured. Similarly, the condition assigned to Subject$_2$ in Wave 1—the No UN Approval condition—differs from the condition assigned to him in Wave 2—the UN Approval condition.

Figure 2 also illustrates the between-subjects design embedded in our experiment. In Wave 1, the condition Subject$_1$ is assigned to—the UN Approval condition—differs from the condition

Subject_2 is assigned to—the No UN Approval condition. However, the probability that Subject_1 (or Subject_2) is randomly assigned to the UN Approval condition (or the No UN Approval condition) in Wave 1 is 0.5—the same for all respondents.

Our experiment differs from existing experiments with within-subjects designs in three ways. First, it is a combined within-subjects and between-subjects design that manipulates the level of a key factor/variable via both within-subjects and between-subjects assignments. This sets apart our experiment from other single-module within-subjects experiments, most of which can be classified into two categories: an experiment with a within-subjects assignment of a key factor accompanied by a separate between-subjects assignment of the key factor (Naoi, Shi and Zhu, 2022; Son and Park, 2022); or an experiment with a within-subject assignment of a key factor and with between-subjects assignments of other factors (Kertzer, Renshon and Yarhi-Milo, 2021; Kertzer and Zeitzoff, 2017; Myrick, 2020, 2023; Renshon, Yarhi-Milo and Kertzer, 2022; Yarhi-Milo, Kertzer and Renshon, 2018; Yarhi-Milo and Ribar, 2022).[14] Both categories of single-module within-subjects experiments focus on testing IR-related hypotheses about treatment effects and utilize within-subjects designs (combined with between-subjects designs) as a tool for doing so (e.g. Bush and Prather, 2021; Kertzer, Renshon and Yarhi-Milo, 2021; Kertzer and Zeitzoff, 2017; Myrick, 2020, 2023; Tingley and Walter, 2011; Yarhi-Milo, Kertzer and Renshon, 2018; Yarhi-Milo and Ribar, 2022). Instead, our experiment focuses on detecting design-induced effects, rather than testing substantial hypotheses. While our experiment features international institutional authorization about a war against a country engaged in nuclear weapons proliferation as the key exogenous variable, this research note does not intend to test theoretical hypotheses about its effect on public opinion.

In other words, most of the existing studies with within-subjects treatments are mixed within- and between-subjects designs in which some factors are assigned within-subjects and other factors

[14] C.f. Bush and Prather (2021) is an exception.

between-subjects.[15] Our design is also a mixed design but the same factor is assigned both within-subjects and between-subjects in order to compare the treatment effects estimated in the within-subjects design and the equivalent in the between-subjects design.

Our set-up is somewhat similar to Bush and Prather (2021)'s design in that one wave is an embedded between-subject design. However, it differs in that the between-subjects design is embedded in our Wave 1 (and, by extension, in Wave 2). In contrast, a between-subject design is embedded in Wave 2 in Bush and Prather (2021)'s study; all subjects in Bush and Prather (2021)'s study were assigned to the same control condition in Wave 1.

Second, our design is optimal for examining design-induced effects across diverse contexts. To assess possible design effects on the cross-country generalizability, our research design resembles the "design of purposive variations" discussed in the recent literature on the generalizability of experimental results across countries and contexts (Bassan-Nygate et al., 2023; Egami and Hartman, 2023). The "design of purposive variations" refers to the experimental design that incorporates the "variations in relevant external validity dimensions" by including "diverse populations, multiple treatments, outcomes, and contexts" (Egami and Hartman, 2023, 1080). This design allows researchers to do "sign-generalization," an assessment of whether the positive or negative sign of the estimated causal effects is generalizable to other contexts, including different outcome measurements and different countries of experimental implementation.

Our decision to field the experiment with different designs in Brazil, China, Japan, and Sweden—a diverse set of countries—allows us to examine the methodological and substantive generalizability of our results. In particular, our cases of four non-Anglo countries are selected to expand the scope of experimental research on IO endorsements, focused on the US and UK. At the same time, they represent a diverse set of attributes that should be relevant to national decisions on the use of force and serve as a moderator of the IOs' effects. All are relatively capable nations, with militaries and economies that can project power. They include an autocracy, a "third wave"

[15] C.f. Bush and Prather (2021) is an exception.

democracy, and two non-Anglo, mature democracies. The four countries differ from each other and from the US and the UK in their history, alliance structures, political alignments, and military postures, important potential moderators of IO effects on public opinion.

We collected a total of 4,214 responses from Brazil; 5,744 responses from one sample in China; 1,866 responses from Sweden; and 888 responses from Japan. Subjects were recruited by professional polling companies in Brazil, China, and Japan and by the University of Gothenburg's Laboratory of Opinion Research for Sweden. The samples are comparable; in all samples, the average respondent was well-educated, middle-class, and displayed some interest in international news.[16]

5.2 A Meta-Analysis of Counterbalanced Within-Subjects Designs

We complement our experiment with a meta-analysis of possible order effects in existing IR studies with within-subjects designs. A meta-analysis is a method that "combines estimates from studies conducted on different samples, in different contexts, or at different times" (Slough and Tyson, 2023, 440) by pooling effect sizes across a literature (Borenstein et al., 2021) and is growing popular in political science (e.g. Bassan-Nygate et al., 2023; Blair, Christensen and Rudkin, 2021; Blair, Coppock and Moor, 2020; Kertzer, 2022; Li, Owen and Mitchell, 2018). Its growing popularity stems from its ability to provide "a systematic synthesis of previous studies" on a particular research question (Jackson and Philips, 2024, 70).

Our meta-analysis focuses on detecting order effects among experiments with the third type of within-subjects designs common in IR research—the repeated-wave, within-subjects design.[17] Specifically, our meta-analysis analyzes experiments that:[18]

[16] See Appendix D.4 for more details on the samples.

[17] See Section 3 for the discussion on types of within-subjects designs common in IR research.

[18] Additional criteria restrict the pool to the IR studies that have publicly available data with the variable for the order of within-subjects conditions for each respondentand are published in the following journals: *American Political Science Review*, *American Journal of Political Science*, *Journal of Politics*, *International Organization*, *International Studies Quarterly*, *Journal of Conflict Resolution*, *Journal of Peace Research*, or *Journal of Experimental Political Science*.

1. employ a within-subjects design—one in which each of the subjects is exposed to all experimental conditions for key factor(s)/variable(s) throughout the experiment;

2. use a repeated-wave design—one in which a respondent is exposed to repeated, successive waves of each of the conditions;

3. randomizes (or counterbalances) the order of the conditions for each subject;

The first criterion excludes the "panel" designs with multiple waves in which: there is no within-subjects assignment of a factor; and between-subjects treatments are implemented to the panel independently in each wave (e.g. Tingley, 2011). In the panel designs, not all subjects are exposed to all conditions because they include those who are assigned to the same conditions over multiple waves.

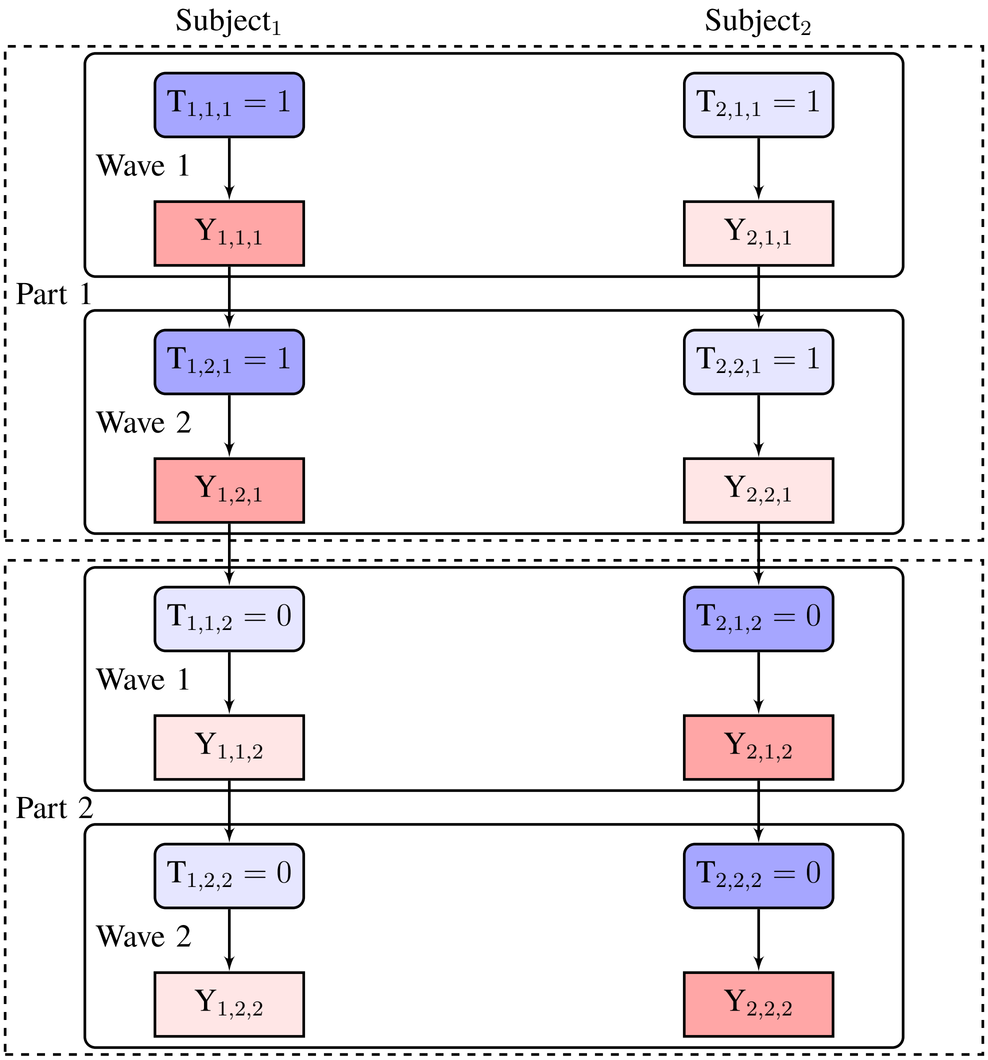
Note that the second and third criteria rule out most of the existing experiments with single-module and multiple-module within-subjects designs (e.g. Böhm, Fleiß and Rybnicek, 2021; Bush and Prather, 2021; Demarest, Jost and Schub, 2024; Naoi, Shi and Zhu, 2022; Son and Park, 2022; Spilker, Nguyen and Bernauer, 2020; Yarhi-Milo, Kertzer and Renshon, 2018; Yarhi-Milo and Ribar, 2022), including our own experiment. We exclude these from the meta-analysis for two reasons. First, our experiment represents these type of within-subjects designs and includes additional features by which within-subjects and between-subjects designs can be compared. It is is analyzed thoroughly in the following sections. Second, it is often difficult to detect order effects from these designs because they were not designed for these purpose. For theoretical reasons, many of the existing studies using these designs first expose subjects to a control condition—to establish a baseline—and then to non-control treatment conditions—often to capture the priming effects (e.g. Bush and Prather, 2021; Naoi, Shi and Zhu, 2022; Son and Park, 2022) or the effects of the variation in another within-subjects factor (Böhm, Fleiß and Rybnicek, 2021). Their treatment effects are measured by comparing subjects' responses to treatments with their responses to the controlled condition as the baseline. This implies that, should order effects exist, it is difficult to disentangle them from treatment effects because all subjects are assigned to treatment conditions after control conditions.[19] Thus, our meta-analysis focuses on detecting order effects in repeated-wave within-subjects design.

We analyze 26 paired conditions from the studies with a counterbalanced within-subjects design.[20] Figure 3 displays an example repeated-wave within-subjects design with counterbalancing. In the example, Subject *i* is assigned to two repeated waves/rounds/plays/games/periods of conditions ($T_{i,1,p}$ and $T_{i,2,p}$ where $T_{i,1,p} = T_{i,2,p}$) in each Part *p* of the two parts. The conditions she is assigned to in Part 2 ($T_{i,1,2}$ and $T_{i,2,2}$ where $T_{i,1,2} = T_{i,2,2}$) differs from those in Part 1 ($T_{i,1,1}$ and $T_{i,2,1}$ where $T_{i,1,1} = T_{i,2,1}$). Due to counterbalancing, the order of conditions each subject is assigned to by part differs by individual. In the example, Subject$_1$ is assigned to Condition 1 in Part 1 and Condition 0 in Part 2 whereas Subject$_2$ is first exposed to Condition 0 in Part 1 and then to Condition 1 in Part 2.

[19] Consequentially, many of these studies complement their within-subjects experiments with either separate between-subjects experiments to bolster their findings (e.g. Naoi, Shi and Zhu, 2022; Son and Park, 2022) or combine them with other between-subjects treatments (e.g. Kertzer, Renshon and Yarhi-Milo, 2021; Renshon, Yarhi-Milo and Kertzer, 2022; Yarhi-Milo, Kertzer and Renshon, 2018; Yarhi-Milo and Ribar, 2022).

[20] See Appendix E.1 for summary statistics.

Figure 3: A Repeated-Wave Within-Subjects Design with Counterbalancing



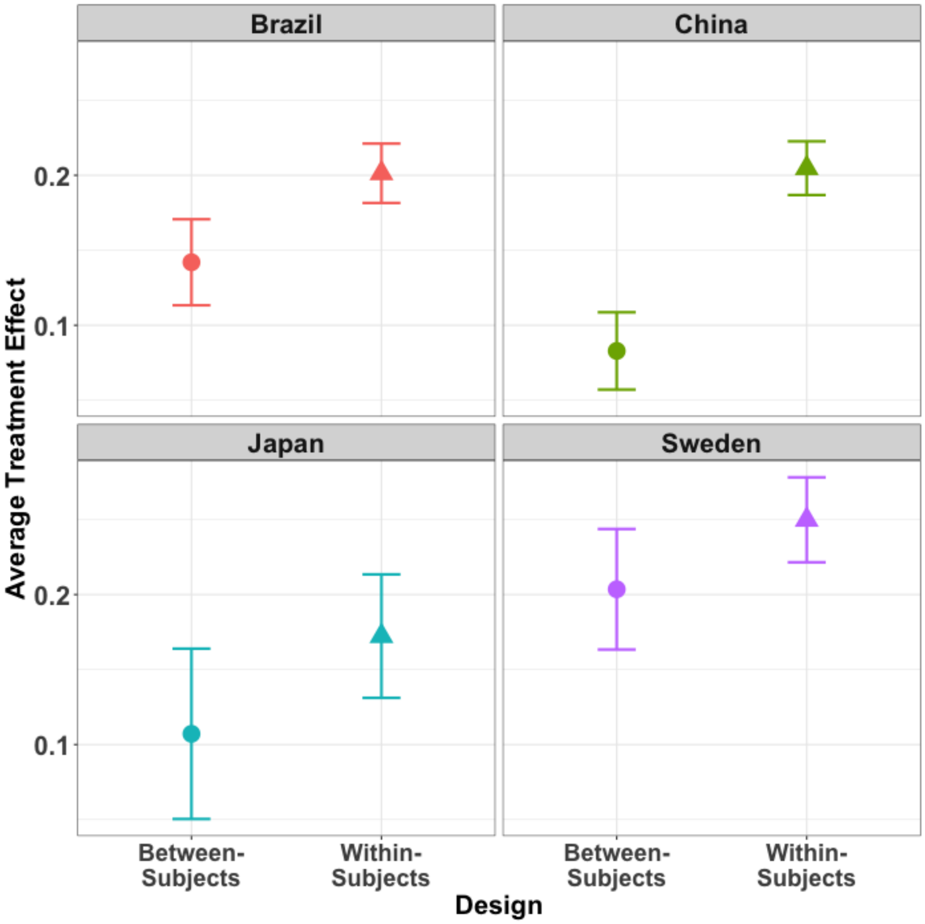
6 Results

We find that within-subjects designs have several tradeoffs.

6.1 Statistical Power With Little Consistency Pressure

First, our within-subjects design generates larger treatment effect estimates with higher precision and lower uncertainty than our post-only, between-subjects design. Figure 4 plots the coefficient estimates of the average treatment effect of UN approval by design.[21]

Figure 4: Average Treatment Effect by Design



Note: The error bars represent the 95% confidence intervals for the coefficient estimates.

The average treatment effect estimates from the within-subjects design are significantly larger than those from the between-subjects design (embedded in Wave 1) for the Brazil and China samples. The former for Brazil is an increase of 20.1% points in support for war and the latter for Brazil is 14.2% points. The Chinese sample's average treatment effect from the within-subject design is an increase of 20.5% points and that from the between-subjects design is 8.29 % points.

The within-subjects design's larger treatment effect estimates suggest that our respondents are

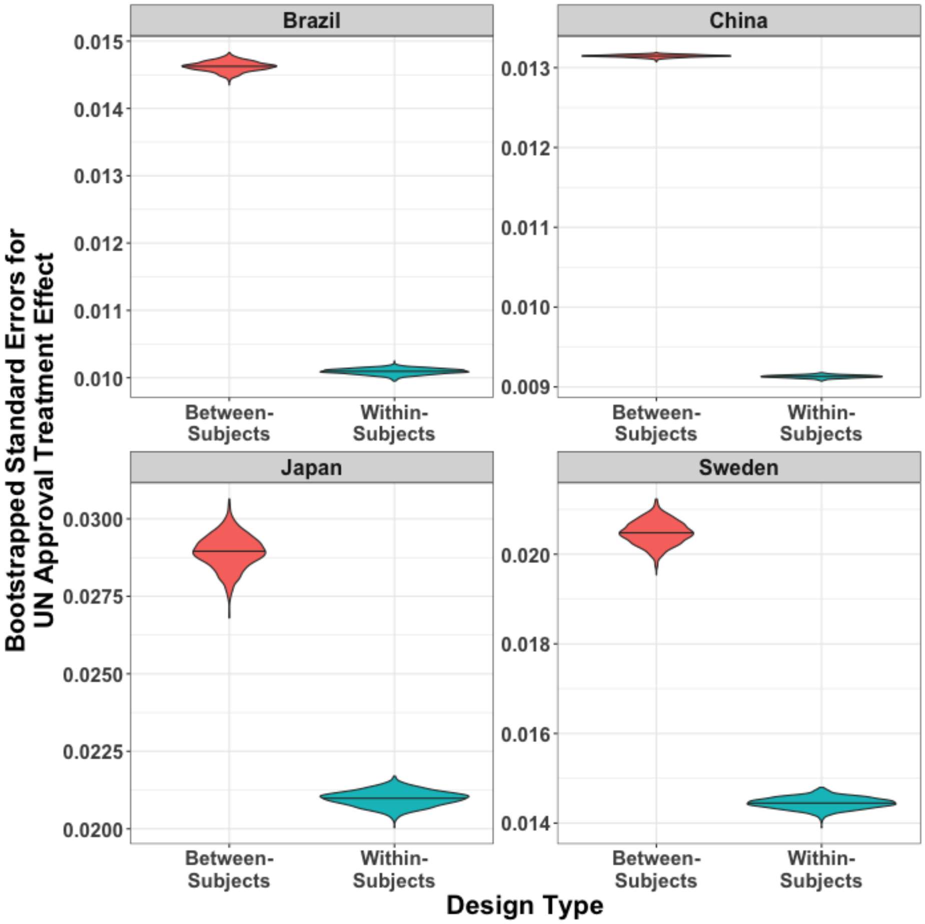
[21] See Tables 17 and 18 for details.

unaffected by the consistency pressure—the urge to stick to their initial answer and ignore the treatment in Wave 2. If it affected them, they would have been less likely to respond to UN approval in Wave 2 than in Wave 1, the embedded between-subjects design, which would have resulted in smaller estimates in the within-subjects design compared to the between-subjects design.

Our bootstrapped standard errors for the estimates of the treatment effects for each design also underscore within-subjects designs' superior statistical power.[22] Figure 5 shows the bootstrapped standard errors.

[22] The treatment effect estimates are equal to the coefficient estimates of UN approval displayed in Appendix D.6's Table 17.

Figure 5: Bootstrapped Standard Errors for the Treatment Effect by Design and Sample



Note: The figure plots the standard error estimates of the treatment effect for UN approval generated from 1,000 bootstrap estimates for each experimental group. The width of each violin corresponds to their relative frequency. The line within each violin represents their median.

For all 4 samples, the bootstrapped standard errors are smaller for the within-subject design than the between-subjects design. For Brazil, the median of the bootstrapped standard errors is 0.010 for the within-subject design and 0.014 for the between-subjects design. Among the Chinese respondents, the median is 0.009 for the former and 0.013 for the latter. For the Japanese sample, the medians are 0.020 for the former and 0.028 for the latter, respectively. For the Swedish sample, the medians are 0.014 for the former and 0.020 for the latter, respectively.

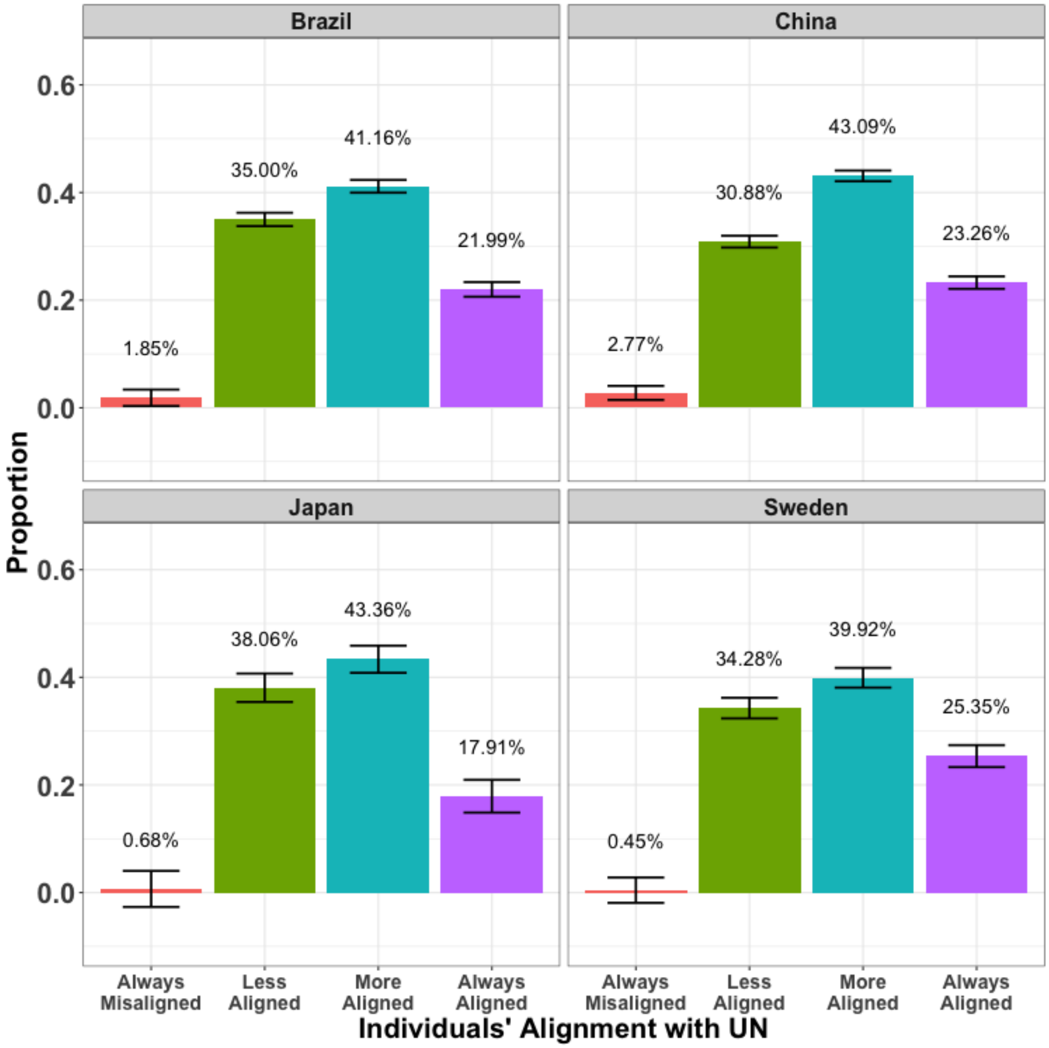
6.2 Heterogeneous Treatment Effects and Demand Effects

However, we find some evidence of demand or learning effects induced by within-subjects designs.

6.2.1 Demand and Learning Effects in Our Within-Subjects Experiment

Our examination of heterogeneous treatment effects in our experiment by Wave implies the existence of the order effects. We examine the heterogeneous treatment effect of IO cues by analyzing each individual's alignment with the UN cue (or lack thereof) by Wave. We classify each sample of respondents to three groups: those whose opinions on the war aligned with the UN in both waves (the "always aligned" group); those whose opinions are consistent with the UN only in one of the two waves (the "partially aligned" group); and those whose opinions were divergent from the UN cue in both waves (the "never aligned" group). Recall that all respondents were exposed to both levels of UN treatments, UN approval in one wave and no UN approval in the other wave, with the order randomized. The "always aligned" respondents are those who supported the use of force when the UN endorsed it in one wave and opposed it in the other wave (when there was no UN endorsement). Respondents who were not sensitive to the UN cues are: The "partially aligned" group are those whose positions in the first wave were divergent from the UN but whose positions in the second wave aligned with the UN in the second wave—the "more aligned (over wave)" respondents—as well as those who agreed with the UN in the first wave but disagreed with it in the second wave—the "less aligned (over wave)" respondents. Figure 6 shows the proportion of respondents by their alignment with the UN cue in each sample.

Figure 6: Individuals' Alignment with the UN Captured by the Within-Subjects Design



Note: The error bars are 95% confidence intervals.

The results underscore the ability of within-subjects designs to capture what between-subjects designs can miss—the partially aligned individuals. Most of our respondents were only partially aligned with IO cues; they did not fully comply with IO cues. The results show that more than 70% of the subjects in each sample followed the cue from the UN only in one of the two waves. This pattern can be captured only by a within-subject design and would have been missed by a between-subject design because the latter classifies the respondents into only two groups—those who are fully aligned with the UN and those who are not. A between-subjects design cannot distinguish the "more aligned" respondents from the "always aligned" respondents or the "less aligned" respondents from the "never aligned" respondents. Substantively, the results imply that the effect of IOs may have been amplified when measured in between-subjects designs and that individuals' loyalty to IOs can fluctuate by context.

However, this analysis suggests the possibility of demand effects induced by within-subjects designs. Note that about 40% of all samples belong to the "more aligned" group, becoming more aligned with the UN in Wave 2 compared to Wave 1 and making this group the largest. If there were no demand effects, most of our respondents would have fallen into the "never aligned" or "always aligned" groups. Instead, the plurality of our respondents seems to grow more responsive to the UN over time.

Moreover, panel linear regression models of public preferences for the use of force in the within-subjects design show that the treatment effect varies by wave in all four samples. Table 2 displays the results of panel linear regression models of individuals' support for the use of force. The dependent variable is coded as 1 for the respondents who support the use of force and 0 for those who do not. The independent variable, UN endorsement, is coded as 1 if the use of force has been approved by the UN and 0 if not. ATEs (average treatment effects) for the within-subjects design are measured in the long-form data with the two waves stacked with respondent random effects, following existing studies (Clifford, Sheagley and Piston, 2021).

Table 2: Panel Linear Regression of War Support in Within-Subjects Design With Random Effects

	Dependent variable:			
	Support for the Use of Force			
	Brazil	China	Japan	Sweden
	(1)	(2)	(3)	(4)
UN Endorsement	0.142^{***}	0.083^{***}	0.107^{***}	0.203^{***}
	(0.014)	(0.013)	(0.030)	(0.020)
Wave 2	-0.095^{***}	-0.179^{***}	-0.006	-0.023
	(0.014)	(0.013)	(0.030)	(0.020)
UN Endorsement*Wave 2	0.118^{***}	0.244^{***}	0.127^{**}	0.094^{***}
	(0.025)	(0.022)	(0.053)	(0.035)
Constant	0.288^{***}	0.474^{***}	0.200^{***}	0.170^{***}
	(0.010)	(0.009)	(0.021)	(0.015)
Observations	8,420	11,483	1,776	3,591
R^2	0.098	0.104	0.100	0.140
Adjusted R^2	0.098	0.103	0.099	0.139
F Statistic	917.791^{***}	$1,325.370^{***}$	197.210^{***}	582.481^{***}
Note:				*p<0.1; **p<0.05; ***p<0.01

The interaction term for UN endorsement and Wave 2 is positive and significant at the 0.05 level for Brazil, China, and Sweden, showing that the average treatment effect for UN approval is larger in Wave 2 compared to Wave 1 for the three samples. Wald tests for these models imply that at least one or more coefficients for Wave-related variables are non-zero for each sample (p-values for Brazil, China, and Japan are 0.000 and Sweden's is 0.002).

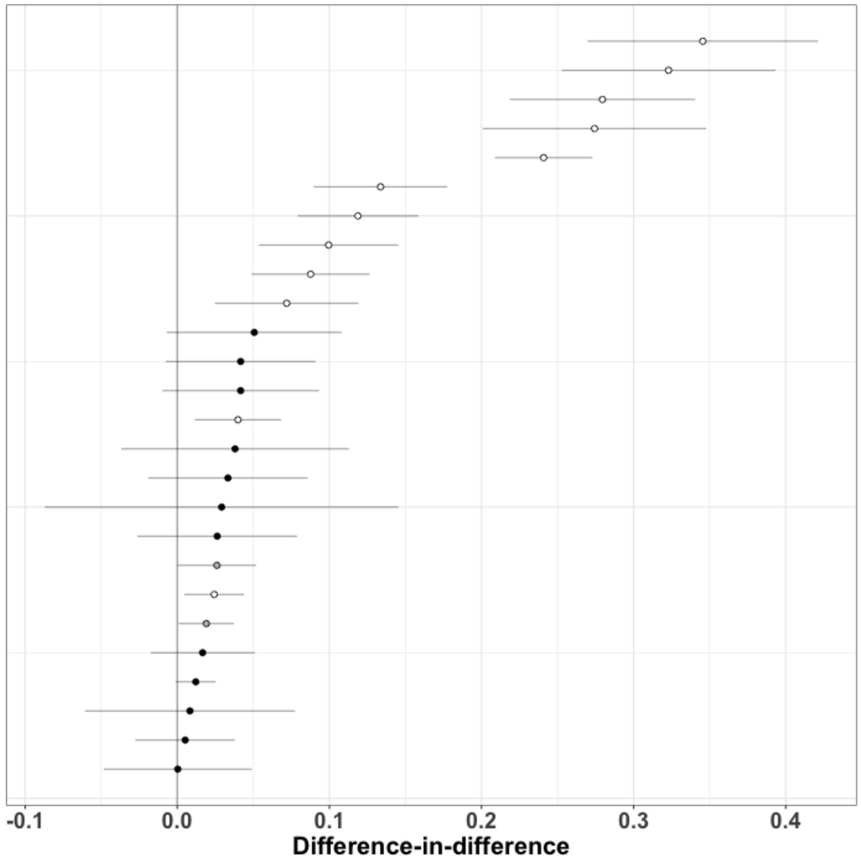
The within-subjects design's larger estimates stem from the larger average treatment effects shown in Wave 2. In Wave 2, the average treatment effect is estimated to be an increase of 26% points in support for the use of force among the Brazil sample and 32.7% points among the Chinese sample. Both estimates from Wave 2 are significantly larger than those from the overall within-subjects design or the between-subjects design.

6.2.2 Order Effects in a Meta-Analysis of Repeated-Wave Within-Subjects Experiments

Order effects also manifest themselves in our meta-analysis of 26 paired within-subjects conditions from lab experiments. To detect the order effects, we estimate the difference-in-differences by the order of conditions applied within-subjects, following (Kertzer, 2022, 543–4). The difference-in-difference estimates measure the absolute difference in a within-subjects factor's treatment effects by the order of conditions of the factor. In other words,

Figure 7 displays the difference-in-difference estimates for each pair of conditions by condition order, with 95% confidence intervals.

Figure 7: Difference-in-Differences by Within-Subject Condition Order across Paired Conditions



Note: The errors bars represent 95% confidence intervals. Estimates statistically significant after controlling for multiple comparisons are shown in white; estimates statistically significant only as long as multiple comparisons are not controlled for are shown in grey.

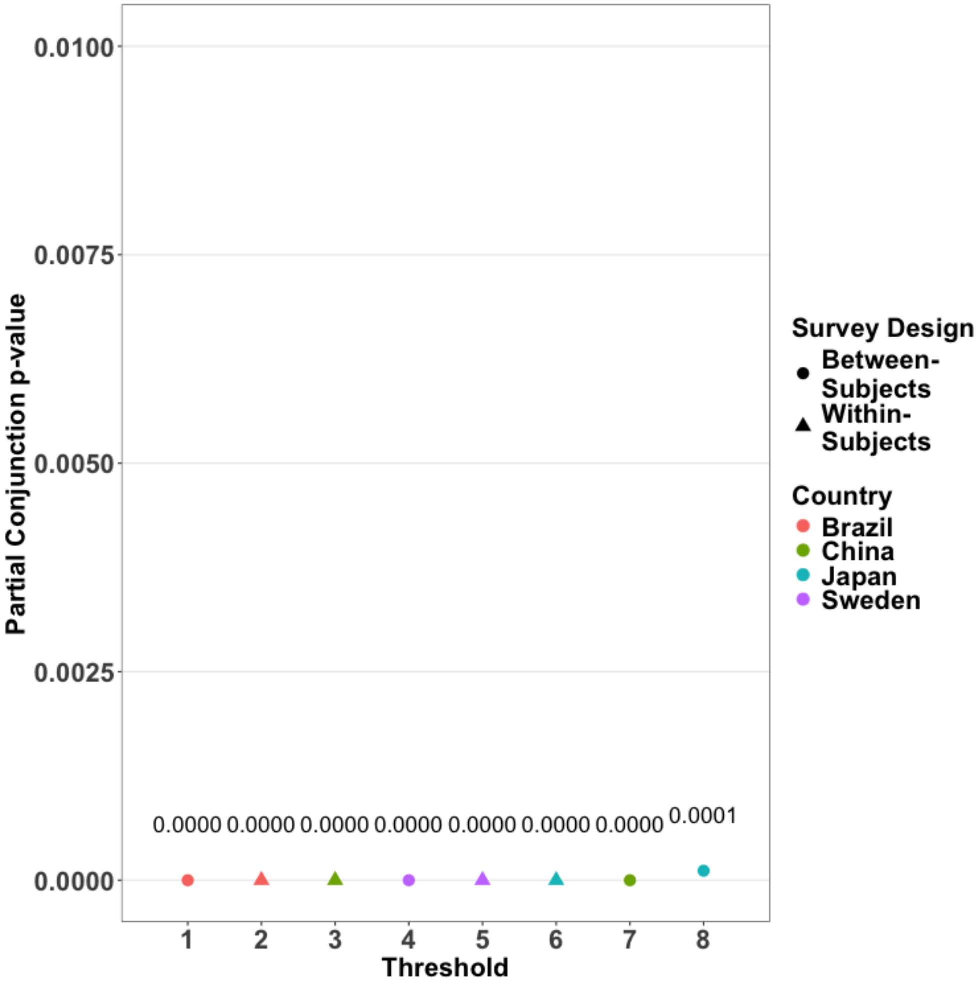
The figure shows that the magnitude of each paired treatment effect (with 95% confidence intervals) can significantly differ by the order of conditions applied within-subjects; the difference-indifference estimates are statistically significant at the p < 0.05 level in 12 of the 26 cases (46.2%), or in 14 of the 26 cases (53.8%) when controlled for the false discovery rate using the Benjamini-Hochberg procedure and following Kertzer (2022).

We check the robustness of the results by checking for interactions between treatments and condition order at the individual-level in Appendix E.2. Again, we find evidence of interactions between average treatment effects and condition order at the individual-level for 2 studies.

6.3 Generalizability

Third, the results of within-subjects designs are generalizable to other contexts. The direction of our treatment effects does not vary by design or by sample. To assess the generalizability of the treatment effects by design or by sample, we run sign-generalization tests (Egami and Hartman, 2023) that utilize our variations in survey designs (within-subjects and between-subjects) and in countries (Brazil, China, Japan, and Sweden). Figure 8 shows the partial conjunction p-values from the sign-generalization tests.

Figure 8: Sign Generalization Test Results



All partial conjunction p-values on both designs and four countries were well below the conventional threshold of 0.05. They suggest that the sign generalizability of our results on IO cues' effect on public opinion is 100%, implying the results are generalizable across different designs and contexts. In other words, the positive treatment effect of UN cues on public support for the use of force did not differ by design or by country.

7 Conclusion

In this paper, we highlight the pros and cons of using a previously uncommon but increasingly popular experimental design—within-subject designs—in IR research (e.g. Kertzer, Renshon and Yarhi-Milo, 2021; Myrick, 2020, 2023; Renshon, Yarhi-Milo and Kertzer, 2022; Tingley and Walter, 2011). We reassess the existing scholarship on their tradeoffs (Clifford, Sheagley and Piston, 2021; McDonald and Hanmer, 2023) and extend it by examining cross-country external validity of their results. We do this using an original survey experiment in Brazil, China, Japan, and Sweden and a meta-analysis of 26 paired within-subjects conditions from existing lab experiments in IR research. We find that within-subjects designs have higher precision, the ability to examine heterogeneous treatment effects, and strong external validity for the direction of the treatment effects, consistent with existing findings (Clifford, Sheagley and Piston, 2021). At the same time, our results suggest the existence of order effects in within-subjects designs, consistent with some existing studies (Chaudoin, Gaines and Livny, 2021; McDonald and Hanmer, 2023; Transue, Lee and Aldrich, 2009) and differing from others Clifford, Sheagley and Piston (2021).

Of course, our study is not without limitations. We note that the designs were not included as a treatment in our experiment; the respondents were not randomly assigned to either a within-subjects design or a between-subjects design. Instead, all respondents were assigned to our within-subjects design in which the first wave was a between-subjects design. Moreover, our within-subjects design allowed no space between the two measures of the outcomes, similar to McDonald and Hanmer

(2023).

Nonetheless, our overall results are meaningful. In particular, our results include not only those from a survey experiment but also those from multiple lab experiments. This extends the discussion on the pros and cons of various experimental designs (Druckman, 2022; Klar, Leeper and Robison, 2020; Montgomery, Nyhan and Torres, 2018; Mummolo and Peterson, 2019; Sheagley and Clifford, 2023), including within-subjects designs (Clifford, Sheagley and Piston, 2021; Chaudoin, Gaines and Livny, 2021; McDonald and Hanmer, 2023) for researchers interested in either survey or lab experiments. Our analysis is particularly meaningful to those less-resourced IR researchers hoping to optimize their resources by providing more information on the options that can lower the barrier to entry to experimental research by their statistical power. Additionally, this paper also proposes an alternative design for the IR experimental literature on the effect of IO cues (Chapman, 2011; Grieco et al., 2011; Matsumura and Tago, 2019), which has been reliant on between-subjects designs.

Future research should investigate the tradeoffs of using conjoint experiments (Tingley, 2014), another experimental design reliant on within-subject comparisons. This research note focuses on standard, non-conjoint within-subjects designs instead of conjoint designs as an alternative to between-subjects designs because the former's accessibility to researchers and the shared estimands of the former and between-subjects designs. However, conjoint experiments are gaining more popularity among experimental IR researchers (e.g. Avey et al., 2022; Kertzer, Renshon and Yarhi-Milo, 2021; Lim and Tanaka, 2022; Majnemer and Meibauer, 2023) and their tradeoffs are worthy of scholarly attention.

References

Adamson, Jordan and Erik O Kimbrough. 2022. "The supply side determinants of territory." *Journal of Peace Research* 60(2):209–225.

Avey, Paul C, Michael C Desch, Eric Parajon, Susan Peterson, Ryan Powers and Michael J Tierney. 2022. "Does social science inform foreign policy? Evidence from a survey of US national security, trade, and development officials." *International Studies Quarterly* 66(1):sqab057.

Bassan-Nygate, Lotem, Jonathan Renshon, Jessica L. P. Weeks and Chagai M. Weiss. 2023. "The Generalizability of IR Experiments Beyond the U.S." APSA Preprints. doi: 10.33774/apsa-2023-dx9kp-v2.

Blackwell, Matthew, Jacob R Brown, Sophie Hill, Kosuke Imai and Teppei Yamamoto. 2023. "Priming bias versus post-treatment bias in experimental designs." *arXiv preprint arXiv:2306.01211* .

Blair, Graeme, Alexander Coppock and Margaret Moor. 2020. "When to Worry about Sensitivity Bias: A Social Reference Theory and Evidence from 30 Years of List Experiments." *American Political Science Review* 114(4):1297–1315.

Blair, Graeme, Darin Christensen and Aaron Rudkin. 2021. "Do Commodity Price Shocks Cause Armed Conflict? A Meta-Analysis of Natural Experiments." *American Political Science Review* 115(2):709–716.

Böhm, Robert, Jürgen Fleiß and Robert Rybnicek. 2021. "On the Stability of Social Preferences in Inter-Group Conflict: A Lab-in-the-Field Panel Study." *Journal of Conflict Resolution* 65(6):1215–1248.

Borenstein, Michael, Larry V Hedges, Julian PT Higgins and Hannah R Rothstein. 2021. *Introduction to meta-analysis*. John Wiley & Sons.

Bowers, Jake. 2011. *Making effects manifest in randomized experiments*. Cambridge University Press pp. 459–480.

Brutger, Ryan, Joshua D. Kertzer, Jonathan Renshon, Dustin Tingley and Chagai M. Weiss. 2023. "Abstraction and Detail in Experimental Design." *American Journal of Political Science* 67(4):979–995.

Bush, Sarah Sunn and Lauren Prather. 2019. "Do electronic devices in face-to-face interviews change survey behavior? Evidence from a developing country." *Research & Politics* 6(2):2053168019844645.

Bush, Sarah Sunn and Lauren Prather. 2021. "Islam, gender segregation, and political engagement: evidence from an experiment in Tunisia." *Political Science Research and Methods* 9(4):728–744.

Chapman, Terrence L. 2011. *Security Approval: Domestic Politics and Multilateral Authorization for War*. Chicago, IL: University of Chicago Press.

Chaudoin, Stephen, Brian J Gaines and Avital Livny. 2021. "Survey design, order effects, and causal mediation analysis." *The Journal of Politics* 83(4):1851–1856.

Chaudoin, Stephen and Jonathan Woon. 2018. "How Hard to Fight? Cross-Player Effects and Strategic Sophistication in an Asymmetric Contest Experiment." *The Journal of Politics* 80(2):585–600.

Chaudoin, Stephen, Sarah Hummel and Yon Soo Park. 2024. "The Election Effect: Democratic Leaders in Inter-Group Conflict." *International Studies Quarterly* 68(1):sqad107.

Cheung, Yin-Wong and Daniel Friedman. 1997. "Individual Learning in Normal Form Games: Some Laboratory Results." *Games and Economic Behavior* 19(1):46–76.

Cheung, Yin-Wong and Daniel Friedman. 1998. "A comparison of learning and replicator dynamics using experimental data." *Journal of Economic Behavior and Organization* 35(3):263–280.

Cialdini, Robert B, Melanie R Trost and Jason T Newsom. 1995. "Preference for consistency: The development of a valid measure and the discovery of surprising behavioral implications." *Journal of personality and social psychology* 69(2):318.

Clifford, Scott, Geoffrey Sheagley and Spencer Piston. 2021. "Increasing precision without altering treatment effects: Repeated measures designs in survey experiments." *American Political Science Review* 115(3):1048–1065.

Crandall, Christian, Owen Cox, Ryan Beasley and Mariya Omelicheva. 2016. "Covert Operations, Wars, Detainee Destinations, and the Psychology of Democratic Peace." *Journal of Conflict Resolution* 62(5):929–956.

Demarest, Heidi, Tyler Jost and Robert Schub. 2024. "Bureaucracy and Cyber Coercion." *International Studies Quarterly* 68(1):sqad103.

Dietrich, Simone, Heidi Hardt and Haley J. Swedlund. 2021. "How to make elite experiments work in International Relations." *European Journal of International Relations* 27(2):596–621.

Druckman, James N. 2022. *Experimental thinking : a primer on social science experiments*. Cambridge, United Kingdom ;: Cambridge University Press.

Egami, Naoki and Erin Hartman. 2023. "Elements of external validity: Framework, design, and analysis." *American Political Science Review* 117(3):1070–1088.

Erev, Ido and Amnon Rapoport. 1990. "Provision of step-level public goods: The sequential contribution mechanism." *Journal of Conflict Resolution* 34(3):401–425.

Fahoum, Nardine, Hadas Pick and Simone Shamay-Tsoory. 2023. "The Impact of Creativity Training on Inter-Group Conflict-Related Emotions." *Journal of Conflict Resolution* p. 00220027231198517.

Gartner, Scott Sigmund. 2008. "The multiple effects of casualties on public support for war: An experimental approach." *American political science review* 102(1):95–106.

Grieco, Joseph M., Christopher Gelpi, Jason Reifler and Peter D. Feaver. 2011. "Let's Get a Second Opinion: International Institutions and American Public Support for War." *International Studies Quarterly* 55:563–583.

Hafner-Burton, Emilie M., Brad L. LeVeck and David G. Victor. 2015. "How Activists Perceive the Utility of International Law." *The Journal of Politics* 78(1):167–180.

Hafner-Burton, Emilie M., Brad L. LeVeck and David G. Victor. 2017. "No False Promises: How the Prospect of Non-Compliance Affects Elite Preferences for International Cooperation." *International Studies Quarterly* 61(1):136–149.

Huff, Connor and Joshua D. Kertzer. 2018. "How the Public Defines Terrorism." *American Journal of Political Science* 62(1):55–71.

Huff, Connor and Robert Schub. 2018. "The intertemporal tradeoff in mobilizing support for war." *International Studies Quarterly* 62(2):396–409.

Hundley, Lindsay. 2019. "The Shadow of the Future and Bargaining Delay: An Experimental Approach." *The Journal of Politics* 82(1):378–383.

Jackson, Christopher and Andrew Q. Philips. 2024. "Synthesize This: Meta-Analysis as a Dissertation Tool." *PS: Political Science & Politics* 57(1):70–75.

Kane, John V. 2024. "More than meets the ITT: A guide for anticipating and investigating nonsignificant results in survey experiments." *Journal of Experimental Political Science* pp. 1–16.

Kertzer, Joshua. 2016. *Resolve in international politics*. Princeton University Press.

Kertzer, Joshua D. 2022. "Re-Assessing Elite-Public Gaps in Political Behavior." *American Journal of Political Science* 66(3):539–553.

Kertzer, Joshua D. and Jonathan Renshon. 2022. "Experiments and Surveys on Political Elites." *Annual Review of Political Science* 25(1):529–550.

Kertzer, Joshua D, Jonathan Renshon and Keren Yarhi-Milo. 2021. "How Do Observers Assess Resolve?" *British Journal of Political Science* 51(1):308–330.

Kertzer, Joshua D. and Thomas Zeitzoff. 2017. "A Bottom-Up Theory of Public Opinion about Foreign Policy." *American Journal of Political Science* 61(3):543–558.

Klar, Samara, Thomas Leeper and Joshua Robison. 2020. "Studying Identities with Experiments: Weighing the Risk of Posttreatment Bias Against Priming Effects." *Journal of Experimental Political Science* 7(1):56–60.

Li, Quan, Erica Owen and Austin Mitchell. 2018. "Why Do Democracies Attract More or Less Foreign Direct Investment? A Meta-Regression Analysis." *International Studies Quarterly* 62.

Lim, Sijeong and Seiki Tanaka. 2022. "Why costly rivalry disputes persist: A paired conjoint experiment in Japan and South Korea." *International Studies Quarterly* 66(4):sqac063.

Lupton, Danielle L and Clayton Webb. 2022. Experimental Methods. In *Routledge Handbook of Foreign Policy Analysis Methods*. Routledge pp. 338–353.

Majnemer, Jacklyn and Gustav Meibauer. 2023. "Names from nowhere? Fictitious country names in survey vignettes affect experimental results." *International Studies Quarterly* 67(1):sqac081.

Martinsson, J., M. Andreasson, E. Markstedt and K. Riedel. 2013. Technical Report Citizen Panel 8 - 2013. Technical report University of Gothenburg LORE.

Matsumura, Naoko and Atsushi Tago. 2019. "Negative surprise in UN Security Council authorization: Do the UK and French vetoes influence the general public's support of US military action?" *Journal of Peace Research* 56(3):395–409.

McDermott, Rose. 2002. "Experimental methods in political science." *Annual Review of Political Science* 5(1):31–61.

McDonald, Jared and Michael J Hanmer. 2023. "Evaluating methods for examining the relative persuasiveness of policy arguments." *Political Science Research and Methods* pp. 1–8.

Mintz, Alex, Yi Yang and Rose McDermott. 2011. "Experimental Approaches to International Relations." *International Studies Quarterly* 55(2):493–501.

Montgomery, Jacob M., Brendan Nyhan and Michelle Torres. 2018. "How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It." *American Journal of Political Science* 62(3):760–775.

Moxnes, Erling and Eline van der Heijden. 2003. "The Effect of Leadership in a Public Bad Experiment." *Journal of Conflict Resolution* 47(6):773–795.

Muller, Laurent, Martin Sefton, Richard Steinberg and Lise Vesterlund. 2008. "Strategic behavior and learning in repeated voluntary contribution experiments." *Journal of Economic Behavior and Organization* 67(3):782–793.

Mummolo, Jonathan and Erik Peterson. 2019. "Demand effects in survey experiments: An empirical assessment." *American Political Science Review* 113(2):517–529.

Mutz, Diana C. 2011. *Population-based survey experiments*. Princeton University Press.

Mutz, Diana C, James N. Druckman and Donald P. Green. 2021. *Advances in Experimental Political Science*. Cambridge University Press Cambridge chapter Improving experimental treatments in political science, pp. 219–238.

Myrick, Rachel. 2020. "Why So Secretive? Unpacking Public Attitudes toward Secrecy and Success in US Foreign Policy." *The Journal of Politics* 82(3):828–843.

Myrick, Rachel. 2023. "Public Reactions to Secret Negotiations in International Politics." *Journal of Conflict Resolution* 68(4):703–729.

Naoi, Megumi, Weiyi Shi and Boliang Zhu. 2022. ""Yes-Man" Firms: Government Campaign and Policy Positioning of Businesses in China." *International Studies Quarterly* 66(4):sqac075.

Ostrom, Elinor, James Walker and Roy Gardner. 1992. "Covenants with and without a Sword: Self-Governance Is Possible." *American Political Science Review* 86(2):404–417.

Rathbun, Brian C., Joshua D. Kertzer and Mark Paradis. 2017. "Homo Diplomaticus: Mixed-Method Evidence of Variation in Strategic Rationality." *International Organization* 71(S1):S33–S60.

Renshon, Jonathan, Keren Yarhi-Milo and Joshua D. Kertzer. 2022. "Democratic Reputations in Crises and War." *The Journal of Politics* 85(1):1–18.

Sheagley, Geoffrey and Scott Clifford. 2023. "No evidence that measuring moderators alters treatment effects." *American Journal of Political Science* .

Slonim, Robert and Alvin E. Roth. 1998. "Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic." *Econometrica* 66(3):569–596.

Slough, Tara and Scott A. Tyson. 2023. "External Validity and Meta-Analysis." *American Journal of Political Science* 67(2):440–455.

Slusher, E. Allen, Gerald L. Rose and Kenneth J. Roering. 1978. "Commitment to Future Interaction and Relative Power Under Conditions of Interdependence." *Journal of Conflict Resolution* 22(2):282–298.

Son, Sangyong and Jong Hee Park. 2022. "Nonproliferation Information and Attitude Change: Evidence From South Korea." *Journal of Conflict Resolution* 67(6):1095–1127.

Spilker, Gabriele, Quynh Nguyen and Thomas Bernauer. 2020. "Trading Arguments: Opinion Updating in the Context of International Trade Agreements." *International Studies Quarterly* 64(4):929–938.

Suong, Clara H., Scott Desposato and Erik Gartzke. 2023. "Thinking Generically and Specifically in International Relations Survey Experiments." *Research and Politics* 10(2):1–6.

Tingley, Dustin. 2011. "The Dark Side of the Future: An Experimental Test of Commitment Problems in Bargaining." *International Studies Quarterly* 55:521–544.

Tingley, Dustin. 2014. "Survey Research in International Political Economy: Motivations, Designs, Methods." *International Interactions* 40(3):443–451.

Tingley, Dustin and Barbara Walter. 2011. "Can Cheap Talk Deter? An Experimental Analysis." *Journal of Conflict Resolution* 55(6):994–1018.

Tomz, Michael and Jessica Weeks. 2013. "Public Opinion and the Democratic Peace." *American Political Science Review* 107(4).

Touchton, Michael R., Casey A. Klofstad, Jonathan P. West and Joseph E. Uscinski. 2020. "Whistle-blowing or leaking? Public opinion toward Assange, Manning, and Snowden." *Research & Politics* 7(1):2053168020904582.

Transue, John E., Daniel J. Lee and John H. Aldrich. 2009. "Treatment Spillover Effects across Survey Experiments." *Political Analysis* 17(2):143–161.

Wilson, Warner. 1969. "Cooperation and the cooperativeness of the other player." *Journal of Conflict Resolution* 13(1):110–117.

Yarhi-Milo, Keren and David T. Ribar. 2022. "Who Punishes Leaders for Lying About the Use of Force? Evaluating The Microfoundations of Domestic Deception Costs." *Journal of Conflict Resolution* 67(4):559–586.

Yarhi-Milo, Keren, Joshua D. Kertzer and Jonathan Renshon. 2018. "Tying Hands, Sinking Costs, and Leader Attributes." *Journal of Conflict Resolution* 62(10):2150–2179.