Annual Review of Political Science

Some Advances in the Design of Survey Experiments

Paul M. Sniderman

Department of Political Science, Stanford University, Stanford, California 94305, USA; email: paulms@stanford.edu

Annu. Rev. Political Sci. 2018. 21:259–75

The *Annual Review of Political Science* is online at polisci.annualreviews.org

https://doi.org/10.1146/annurev-polisci-042716-115726

Keywords

cross-category comparisons, null by design, conjoint, sequential factorials, splicing, replication

Abstract

This article calls attention to some designs in survey experiments that give new leverage in hypothesis testing and validation. The premise of this review is the modesty of survey experiments—modesty of treatment, modesty of scale, modesty of measurement. The focus of this review, accordingly, is the compensating virtues of modesty. With respect to hypothesis testing, I spotlight (*a*) cross-category comparisons, (*b*) null-by-design experiments, (*c*) explication, (*d*) conjoint designs, and (*e*) sequential factorials. With respect to validation regimes, I discuss (*a*) parallel studies, (*b*) paired designs, and (*c*) splicing. Throughout, the emphasis is on moving from experiment in the singular to experiments in the plural, learning as you go.

INTRODUCTION

> In Athens and Ionia the Greeks invented what may be called First Science. . .whose method consisted in the postulation of axioms from which certain theorems can be deduced from the application of a logical system of inference. . . . It was not until nearly a score of centuries later that First Science gave way to Second Science, the product of the Renaissance, which systematized the procedure of experimentation.
>
> —Conrad Waddington (1970, pp. 1–2)

My aim in this review is to call attention to designs in survey experiments that give new leverage in hypothesis testing and validation. My particular concerns: design rather than analysis; survey experiments rather than experiments generically; and, notwithstanding the special attention paid to replication and reproducibility, bringing into view what we have not yet seen rather than demonstrating that we indeed know what we already had good reason to believe that we knew. The largest architectural lesson to draw from advances in the design of survey experiments: conceiving public opinion studies not as one-off omnibus surveys but as a progression of experimental trials, simultaneously extending and cross-validating lines of research.

The efficiencies of survey experiments are my theme. The first note I therefore must strike is their limitations.

MODESTY

Survey experiments combine representative samples and randomized assignment—surely a combination that makes for rigorous science. But think of what a survey experiment characteristically consists of—a single question, appearing in multiple versions to be sure, but almost certainly novel in content, which is to say, without a track record. Moreover, the term treatment, although technically justified, teeters on the edge of grandiosity, characteristically denoting a few variations of phrases or paragraphs read to or read by respondents. Randomly varying whether black individuals or white individuals benefit from a job training program is not even the ninth cousin twice removed of injecting an experimental Ebola vaccine.

One reason for the humbleness of survey experiment interventions is high-minded ethical constraints. Another—and, as I mean to spotlight, neglected—consideration is the conversational constraints of a public opinion interview.[1] All in all, it is closer to the mark to think of a treatment in a survey experiment not as an intervention or manipulation, but as a variation in information presented or highlighted for respondents.

Modesty of treatments is one limitation of survey experiments. Another is modesty of scale, especially duration—or, more accurately, the lack of duration. Most often, survey experiments are over almost as soon as they start. And not the least form of modesty is modesty of measurement. Most often, only a single dependent variable is measured, and, what is more, it is measured by only a single indicator. It is not logically necessary for this to be the case. Nevertheless, it is all too frequently the case.

These three limitations of survey experiments—modesty of treatment, modesty of scale, and modesty of measurement—need constantly to be borne in mind when brandishing the term experiment as a prestige enhancer.

[1]Ease of reading and comprehension is a necessity for general population surveys, representative or otherwise. Simplicity of expression is, accordingly, the order of the day for complex as well as simple ideas. The implications for ecological validity and generalizability are worth consideration.

EXPERIMENTAL DESIGN AND HYPOTHESIS TESTING

Background

The primary focus of the first generation of survey experiments was issue framing (Gamson & Modigliani 1989). A frame, in this context, "is a central organizing idea" (Gamson & Modigliani 1989, p. 3). Frames, as Nelson & Kinder (1996, p. 1057) observe, make salient "how it should be thought about, and may go so far as to recommend what (if anything) should be done."

A prototypical example of a framing experiment is Nelson et al.'s (1997) study of issue framing and support for civil liberties. Creating videotapes mimicking local television newscasts, Nelson et al. (1997) presented respondents in one experimental condition with a story highlighting the value of free speech and respondents in a second condition with a story highlighting the risk of violence and disorder. Comparing the two experimental conditions, they found that, when the story was framed in the first way, participants' support for freedom of speech was strong; when the story was framed in the second way, participants' support was weak.

This discovery of preference reversals ignited the study of framing effects. The elasticity of the concept of framing has been a mixed blessing. Clarity requires discipline. Druckman (e.g., 2001) has taken the lead in developing a specifically political framework. Together with his colleagues, he has systematically drawn out and investigated the implications of a theory of issue competition (see especially Chong & Druckman 2007, 2010; Druckman et al. 2013). A natural extension has been investigation of political parties and issue sponsorship. Slothuus and his colleagues (e.g., Leeper & Slothuus 2016, Slothuus 2010) have broken new ground by tying together survey and natural experiments. Sometimes by anticipating events and sometimes by exploiting opportunism as a methodological strategy, they have captured reactions to uncommon but telling political events—for example, the responses of partisans to their party's reversing its position on a salient issue. Their methodological ecumenism deserves particular mention: They analyze the interplay of elite cues and general predispositions in opinion formation, for example, through a multiwave panel incorporating randomized experiments. A recent offshoot of party-centered studies has been a new perspective on spatial reasoning. Previous studies, consciously or otherwise, have zeroed in on proximity judgments of candidate positions. But electoral choices are double choices—of a candidate and of a party. Survey experiments open the door to simultaneously varying distances both between candidates' and voters' positions and between candidates' positions and the policy reputations of their parties. The effect is to bring out a basis for ideologically grounded judgments by partisans (Sniderman & Stiglitz 2012).

The logic and methodology of framing experiments have been at the center of attention for decades. I concentrate, therefore, on the most recent developments.

Cross-Category Comparisons

An original impulse behind the introduction of survey experiments was to move beyond the measurement of prejudice to the assessment of discrimination (Sniderman & Piazza 1993, Sniderman et al. 1991). The design objective, accordingly, was to identify racially distinct groups (or individuals representative of racially distinct groups), each with a credible claim to government assistance. By randomly varying the group on behalf of which the claim was made, one can determine the extent to which a majority treats a minority differently and worse. The claim may be for assistance from the government (e.g., a job training program), for protection against the power of the state (e.g., freedom of speech), or for protection by the state against societal discrimination (e.g., affirmative action). The idea of research in this area is to determine if a claim made on behalf of similarly situated groups or individuals is honored for whites but not for blacks. Similarly situated, in this context, means equally warranted to claim a right or benefit.

Groups or individuals similarly situated are a necessary condition for discrimination designs. This is so manifestly the right road to take in so many situations that it has obscured another—and for some purposes better—route. Consider the civil liberties experiment designed by Petersen et al. (2011) to assess the reactions of Danes to Muslims during the controversy over publication of cartoons depicting the prophet Mohammed. The objective was to assess support for the civil liberties of Muslims. Respondents were asked a battery of standard questions measuring support for civil liberties for a randomly assigned group. Two of the groups were Islamic fundamentalists and Muslims. Drawing on a nationally representative sample of adult Danes, Petersen et al. found markedly lower levels of support for the former than for the latter.

This result is hardly surprising and is, so far, standard. Support for civil liberties of Muslims is higher than that for Islamic fundamentalists. But how high must support for civil liberties for Muslims be to be considered high? The design of the Petersen et al. experiment facilitated an inspired cross-category comparison—namely, with born-again Christians. Christian fundamentalists are in conflict with mainstream Danish society, protesting abortion, homosexuality, and the secularization of society. But Danish Christians are an undeniably legitimate group in Denmark. Hence the power of the Petersen et al. findings. Danes gave equal support to the civil liberties of Muslims and fellow Danes, and they did so at the peak of the most intense political crisis since World War II, in the face of repeated demands by Muslim countries that respect for Islam take precedence over freedom of speech in Denmark.

The making of multiple—and thus replicable—within- and across-category comparisons is a further feature of the Petersen et al. design that is worth attention. The overall hypothesis is that Danes draw a categorical distinction between Islamic fundamentalists and Muslims—a fact that is important in its own right, if indeed it is a fact. Reactions to eight groups in all were assessed: four associated with violent acts, four not. A test of whether citizens draw a categorical distinction between Islamic fundamentalists and Muslims is similarity of responses to Islamic fundamentalists to each of three violent groups plus dissimilarity to each of three nonviolent groups. Ditto but the other way around for Muslims.

A second illustration of the power of cross-category experiments is the right to assembly experiment, designed by Ivarsflaten and her colleagues (Ivarsflaten & Sniderman 2017) and conducted as part of the Norway Citizen Panel (Berntzen et al. 2017). The objective was to assess the perceived political legitimacy of an array of groups. Like Petersen et al.'s (2011) study, the right to assembly experiment randomized assignment across, and not just within, categories or groups. The test of perceived legitimacy is popular support for renting a local community center to a group to house a meeting of their members and sympathizers. Among the assignments, a subset of respondents was assigned a Christian congregation, another a Muslim congregation.

To learn that a national majority is just as willing to rent the local community center to a Muslim as to a Christian congregation is genuinely instructive. But the use of cross- rather than within-category comparisons opens the door to a more striking discovery. Exercising freedom of assembly is part of the mission DNA of a political party. Political parties are thus the gold standard reference point to assess support for civil liberties. It is all the more impressive, therefore, that the level of support for a Muslim congregation renting the local hall for a meeting of its members and sympathizers is—as near as makes no difference—as high as for a political party holding a meeting.

Null by Design

A second design strategy for hypothesis testing is null by design. The hallmark of null-by-design experiments is striking variations in treatments that seem sure-fire bets to evoke differential responses, but—as investigators had predicted—prove to be squibs. Failure, if you will, is success.

Dosage experiments lend themselves to null-by-design experiments. An example is the SAT experiment of Sniderman & Piazza (1993). Their hypothesis was that black Americans are committed to the core values of the larger American culture. To test their commitment black Americans are forced to choose between honoring a core value like achievement and giving a helping hand to a fellow black American. All respondents were black, all were informed about two candidates for admission to a college, and all were told that only one candidate could be admitted. One candidate was black, and the other was white. The white candidate always had the higher score on the admission test (the SAT), 80 out of 100. The gap in scores between the black and white candidates was randomized in increments of 5, from very large (55 points) to very small (5 points). The social background of the candidates was also randomly varied between working class and middle class. The test question was: If only one of the two candidates can be admitted, who should it be? Sniderman & Piazza predicted that, forced to choose between the two candidates, black Americans would choose the white candidate over the black candidate. They would do so because the white candidate had the higher admission test score. And, this is key, they would be equally likely to do so however modest the difference in the SAT scores of the two candidates, and—just to make things more interesting—notwithstanding a reminder that black individuals still have to work harder than white individuals to do as well.

In null-by-design experiments, everything hinges on the variation in experimental conditions seeming a sure bet to evoke differential responses. A panel of experts was assembled independent of the survey (Sniderman & Piazza 1993, p. 147). They were asked at what point the difference in scores between the black and white candidates would be so small that most black respondents would treat it as inconsequential and favor a fellow black candidate for college admission. All judges, without exception, predicted that a majority of African Americans would favor a fellow African American when the difference in test scores between the black and white candidates was tiny, e.g., only five points.

The actual outcome of the SAT experiment was the opposite of the judges' predictions. African Americans overwhelmingly chose the candidate with the higher score. They did so even though this meant turning down the application of a fellow African American and despite being asked to consider the systemic disadvantages that black Americans suffer. Above all, African Americans overwhelmingly turned down a fellow African American however small the difference in test scores. The willingness of African Americans to choose the white candidate over the black candidate, because the white candidate had a higher score on an achievement test, is evidence of the sincerity of black Americans' commitment to the value of achievement.

A second, and particularly ingenious, example of a null-by-design experiment is Grimmer et al.'s (2015) study of Congressional credit claiming. They asked the question: How much do legislators have to accomplish for constituents to win their support? Two features of their design are especially noteworthy. The first has to do with ecological validity. A major method of credit claiming is Congressional press releases announcing work undertaken on behalf of constituents. The screen display used in the experiment—the interviews were online—mimicked a Congressional press release, illustrating the injunction that experimental design should be grounded in substantive savvy.

Grimmer et al.'s hypothesis was that constituents will respond as favorably to a legislator who asks for a benefit for his constituency as to one who actually delivers one. This is not a garden-variety, "no more than common sense" hypothesis. It seems a take-the-money-and-run bet that constituents will respond more favorably to legislators who bring home the bacon than to those who merely say that they will request some bacon to bring home. Their experiment accordingly varies, among other factors, whether legislators report that they have secured, or requested, or will request funds to repair local roads. By design, this information is repeated twice during the

course of the experiment whether the legislator has succeeded in securing, has asked for, or is merely asserting that they will ask for funds to repair local roads.

It is common sense that actually delivering the goods earns legislators more good will than announcing an intention to ask for them, but it is wrong nonetheless. In Grimmer et al.'s experiment, legislators earned as much good will when they did the least amount possible, i.e., saying that they will ask for help for their constituents, as when they accomplished the job, i.e., actually securing the funding.

This is the outcome that Grimmer and his colleagues anticipated. Null-by-design designs are most at risk of going off the rails by failing to detect a between-conditions difference, when in reality responses between conditions differ. So, at the design stage, Grimmer and his colleagues consciously took the precaution of ensuring that they would not fail to reject the null hypothesis for lack of statistical power.

Explication

Prototypically, survey experiments are conducted to determine whether responses in treatment and control conditions are significantly different. But finding that a treatment matters carries you only part way down the road. Once you know a treatment matters, it is necessary to learn why it matters.

Designing survey experiments to explicate, rather than merely to demonstrate, effects requires a depth and exactness of understanding that is uncommon, at any rate in the neighborhoods of social science where I know my way about. Design is all about anticipation, about developing instruments in advance of data collection to conduct particular data analyses. Designing for explication, moreover, invites one to consider multiple competing hypotheses when there is good reason to expect that more than one may be right.

An example of an experiment designed for explication of treatment effects is a deeply thought-through study of the democratic peace by Tomz & Weeks (2013). Democracies almost never fight one another. The question is why. The problem is not coming up with an explanation. The problem is a superabundance of explanations.

Tomz & Weeks isolated three theories to explain why citizens are less likely to support a war against a fellow democracy: (*a*) They are less likely to perceive a fellow democracy as a threat, (*b*) they are more likely to perceive a war with a fellow democracy as entailing high costs and low returns, and (*c*) they are more likely to perceive a war with a fellow democracy as violating considerations of morality. All, some, one, or none of these theories may be valid. The Tomz & Weeks study is an example of how one can design a progression of studies to extend and deepen one's understanding. The study, conducted in the United Kingdom, showed citizens reluctant to go to war with fellow democracies even after taking into account other considerations, e.g., whether an alliance had been formed. The next question to be answered is why.

For leverage, Tomz & Weeks extend their design, specifying alternative causal mechanisms—or mediators. The mediational component, I would emphasize, is an integral part of the design of the Tomz & Weeks experiment, custom engineered to test hypotheses specified in advance.

The mediation analysis showed threat perception to be the primary mediator, with considerations of morality a secondary mediator, an important substantive discovery. But what makes the Tomz & Weeks study standout work is, first, the development of a design to establish not only that a treatment works but also why, and, second, an appreciation that mutually competitive hypotheses need not be mutually exclusive. Appreciating the fact that mutually competitive is not a synonym for mutually exclusive raises the level of play. It is no longer enough for investigators to show that a hypothesis beats the null. They should also show that it beats credible competitors. In

addition, and still more importantly, although methodological expertise is a design requirement, the big lesson of the Tomz & Weeks study is that creativity in design is grounded in a capacity to carve a problem at its analytical joints.

Conjoint Designs and the Analysis of Multidimensional Choice

Conjoint designs are arguably the most promising design innovation in survey experiments developed over the past decade. A standard survey experiment can vary only a few factors. Conjoint designs can vary an indefinite number of factors, each one orthogonal to every other, all capable of assuming multiple values.

Comparative advantages of conjoint designs. Conjoint designs have a wide range of applications. An issue like immigration showcases their strengths in assessing the validity of rival explanations. According to one line of explanation, native citizens object to immigration and immigrants because they anticipate either that they personally will be materially worse off because of stiffer labor market competition or higher taxes (e.g., Scheve & Slaughter 2001) or that the country as a whole will be worse off because of recession and unemployment (Citrin et al. 1997). According to a competing line of analysis, native citizens object to immigration because they anticipate a failure of immigrants to fit in culturally, socially (Hainmueller & Hopkins 2014, Sniderman & Hagendoorn 2007), or both (e.g., Bansak et al. 2016).

Which provides the better explanation, a political economy perspective or a social psychological one? To answer this question, three obstacles have to be overcome. First, some attributes are systematically ambiguous. Is the amount of education that immigrants have achieved an indicator of their skill level, in the eyes of native citizens (Scheve & Slaughter 2001), or a signal of their cultural values (Ford et al. 2012, Hainmueller & Hiscox 2010)? Second, choices about immigration and immigrants are multidimensional. Possibly relevant factors, to tick off those that come quickest to mind, are the immigrant's previous familiarity with the destination country, reason for application, country of origin, language skills, profession, work experience, level of education, and gender. Third, the presentational form of standard survey experiments is sentences and paragraphs. Both can be cumbersome, limiting the number of attributes that can be varied, not to mention the degree of specificity with which they can be characterized. In comparison, a conjoint design allows an indefinite number of attributes to be economically characterized. Equally advantageous, point-to-point comparison of attributes facilitates comparative evaluation. The result: the pattern of evaluation of immigrants' attributes reduces the potential ambiguity of each attribute and provides a multifaceted assessment of competing explanatory perspectives.

What Is New This Time Around?

It may be asked why conjoint designs should be described as an innovation. The same technique, under the name of vignette analysis, was introduced decades ago (e.g., Alves & Rossi 1978, Rossi et al. 1974). Between then and now, conjoint analysis has led a thriving life in consumer research (Green & Srinivasan 1990). But if it is reasonable to ask what distinguishes the practice of conjoint analysis this time around, the answer is easy enough to give.

It is the establishment of the methodological foundations of conjoint analysis by Hainmueller and his colleagues. Their measurement program, undertaken in parallel with their substantive studies, has been detailed, systematic, and creative. It includes analyses of the upper limit of attributes to be evaluated (Bansak et al. 2017); the number of choice tasks to be performed (Bansak et al. 2018); causal inference in conjoint analysis (Hainmueller et al. 2014); and, not the least accomplishment, the validation of conjoint analyses of immigrant approval judgments against actual decisions (Hainmueller et al. 2015).

Conjoint analysis has been criticized for opening the door too widely in hypothesis testing. It legitimates, runs the argument, the simultaneous test of all possibly plausible explanations. In the case of immigration, the dimensions of choice—e.g., the immigrant's occupation, facility in the host country's language (or lack thereof)—are chosen because a reason for predicting their relevance has been established. But there is no ex ante, foot-in-cement prediction which of the competing hypotheses will fare better and which will fare worse. To this extent, conjoint experiments do not satisfy strict conceptions of preregistration. So much the worse for preregistration one might reply. If we already knew whether economic self-interest or social-psychological factors dominate preferences over immigrants, an experiment would be unnecessary. The whole point of performing an experiment is that we do not already know the answer.

Sequential Factorials

The modesty of survey experiments is the opening theme of this review. But modesty has compensating virtues, as any fan of Jane Austen knows. Survey experiments invert the economics of survey research. The cost of interviews is considerable. The cost of one question is tiny. This cost differential allows one to move from a model of an experiment to test a hypothesis to a model of a sequence of experiments to progressively deepen and draw out the implications of a line of reasoning. This can be accomplished in a variety of ways, but one design with special advantages is the sequential factorial.

An example: suspect loyalty. By a sequential factorial, I mean a series of experimental trials that hold constant the design template but progressively vary the values of the factor(s) being manipulated to refine and deepen a line of inquiry. As an example, consider Islamophobia. Muslim immigrants are the target of stereotypes applied to other minorities, including overwillingness to exploit welfare benefits, to complain about their problems and try to make others feel sorry for them, and to avoid doing hard work. However, Muslims are also specially branded, most consequentially, as politically untrustworthy and more loyal to the country they came from than to the country they now live in (Sniderman & Hagendoorn 2007).

Research on Islamophobia has focused on hostility toward Muslims, that is, ordinary Muslims. But the stigma of disloyalty suggests a potential hypothesis: If the loyalty of ordinary Muslims is suspect, is it not likely that the loyalty of the leadership of Muslim communities is even more so? If so, Muslim leaders will be doubly suspect: first, because they are Muslims and second, because they are leaders of a Muslim community.

This conjecture has no provenance: It is not deduced from a theory; it is an intuition, a shot in the dark, if you will. I say this by way of acknowledgment, not by way of apology. **Table 1** reports a series of experimental trials, an example of a sequential factorial, holding the operational design constant while progressively varying the values of the factor being manipulated to deepen and to refine a line of inquiry. Holding the operational design constant facilitates replication across time and, in this case, countries.[2]

The layout of **Table 1** reports, column by column, a sequence of studies, beginning in the Netherlands and then alternating between Denmark and Norway (Ivarsflaten & Sniderman 2017). The numerals accompanying the names of the countries identify the order of the study in each country (e.g., Denmark 3 is the third survey conducted in Denmark as part of the overall study; Norway 8 the eighth survey conducted in Norway as part of the overall study). The number of

[2]The research was supported by the National Science Foundation EAGER Exploratory Grant SES-1449162.

Table 1 Suspect loyalties: percentage of respondents who suspect the group of disloyalty (Ivarsflaten & Sniderman 2017)[a]

Group	Experiment 1: Netherlands 2[b]	Experiment 2: Denmark 2[c]	Experiment 3: Denmark 3[c]	Experiment 4: Norway 6[d]	Experiment 5: Norway 8[d]
Muslims	34	46	45	24	30
Muslim leaders	55	60	66	39	48
Moderate Muslims	27	NA	39	16	NA
Muslim religious leaders	NA	64	69	NA	52
Muslim fundamentalists	NA	NA	NA	82	NA
Muslim extremists	NA	NA	NA	NA	87
Local Muslim politicians	NA	NA	NA	NA	24

Abbreviation: NA, not asked.

[a]The numerals accompanying the names of the countries identify the order of the study in each country (e.g., Denmark 3 is the third survey conducted in Denmark as part of the overall study; Norway 8 is the eighth survey conducted in Norway as part of the overall study).

[b]Data source: TNS NIPO. Response format: 6-point Likert.

[c]Data source: YouGov. Response format: 5-point Likert.

[d]Data source: Norwegian Citizen Panel. Response format: 5-point Likert.

experiments and the number of countries should be read as an advertisement for the benefits of the minimal cost of a survey experiment. The rows of **Table 1** report the values of the randomized factor (e.g., for the Netherlands, the referents are Muslims, Muslim leaders, and moderate Muslims; for Denmark 3, the referents are Muslims, Muslim leaders, and Muslim religious leaders.)

The pattern of factorial values, through the progression of experimental trials, summarizes the logic of progressively targeting the specific meaning of Muslim leaders. The starting point is the question: Are Muslim leaders burdened by a double layer of distrust? Representative national samples were asked: "How much do you trust or distrust [Muslims/Muslim leaders/moderate Muslims] when they say that they want to become part of our country?" The quantity of interest was the similarity or dissimilarity of responses to Muslims and Muslim leaders. Moderate Muslims are included as a reference point. The first column of **Table 1** reports the results of an experiment in the Netherlands, the second a partial replication and extension in Denmark, and the third a further replication in Denmark.

The Dutch results make the point: Muslim leaders are markedly more likely than Muslims to be suspected of disloyalty (55% compared to 34%), while moderate Muslims and Muslims are perceived quite similarly (27% compared to 34%). The pattern of results of the partial replication and extension shown in column 2 and of the complete replication shown in column 3 (both conducted in Denmark) is identical. The specific figures are not. Heavy breathing over between-country differences would be a mistake. Part of the concern over reproducibility is self-flattery about exactness of measurement. Measures of opinion are ordinal, assumptions of equal intervals unwarranted, and claims beyond "more" or "less" unjustified.

Ivarsflaten & Sniderman's research program has made a principled commitment to coarse measurement—for example, ex ante acceptance of different survey organizations' different response formats. The premise: robust patterns, not point estimates, are a realistic target. In every country, the pattern within and across countries is the same. There is a striking dissimilarity between responses to Muslim leaders, on the one hand, and Muslims and moderate Muslims on the other.

The obvious question to ask is: Who do national majorities have in mind when they think of Muslim leaders? The strategy for answering this is a search for synonymy. Terms are synonymous

just so far as they are substitutable sine mutandis. Accordingly, a second stage variation of the Suspect Loyalties template assesses responses to Muslim leaders and Muslim religious leaders, as well as Muslims as a reference point for both. The results, shown in columns 2 and 3 of **Table 1**, show the loyalty of Muslim leaders and Muslim religious leaders to be similarly suspect.

Muslim leaders, this suggests, are interchangeable with Muslim religious leaders in the minds of majority citizens, which points to the next question to ask: Who (or what) do majority citizens have in mind when they think of Muslim religious leaders? The most obvious possibility is religious fundamentalists. Accordingly, the third step in this sequence of experiments is to assess responses to Muslim fundamentalists (while replicating previous results). As the results of the fourth study, conducted in Norway, show, Muslim leaders and Muslim fundamentalists are not interchangeable: There is markedly more distrust of the latter than of the former. As a follow-up, the fifth study, again conducted in Norway, assesses the similarity of responses to Muslim religious leaders and to Muslim extremists (while replicating the core result of greater suspicion of Muslim leaders than of Muslims). Again, far more distrust the latter than the former.

Muslim leaders are thus doubly suspect: First, because they are Muslims, and second, because in the minds of majority citizens, "Muslim leader" is interchangeable with "Muslim religious leader"—but "Muslim religious leader" is not a synonym for "Muslim fundamentalist" or "extremist." It is a happy outcome that "Muslim religious leader" and "Muslim fundamentalist" are not interchangeable. But the fact remains that far more distrust Muslim leaders, because they are taken to be Muslim religious leaders, than distrust ordinary Muslims. It is obviously necessary to ask one more question: How easy is it to inhibit majority citizens thinking "Muslim religious leader" when they see or learn about a Muslim leader?

The final variation of the Suspect Loyalties template points to an answer. The stronger is the automaticity of the association of Muslim leaders with Muslim religious leaders, the more similar is the response to other characterizations of Muslim leaders. A test, then, is the similarity of responses to Muslim local politicians. The results, displayed in the sixth column of **Table 1**, show that the connection between Muslim leaders and Muslim religious leaders is relatively easily inhibited.

This is a one-hand-tied-behind-the-back presentation of substantive results, since my concern is experimental design. All the same, the sequence of Suspect Loyalties experiments provides a lesson in step-by-step examination of synonymy. The larger lesson: learn as you go.

Cultural pluralism. It is widely agreed that inclusion of minorities requires a willingness to acknowledge the worth of their culture and values (Crowder 2013, p. 7). Are national majorities willing so? To ask this question is to ask a bundle of questions. What does it mean, in the minds of majority citizens, to acknowledge or recognize the worth of a minority's culture and values? How far are national majorities willing to go? Where do they draw the line? Why do they draw it there and not elsewhere?

A first step in attempting to answer these questions is to look for lower and upper bounds on the readiness of national majorities to approve of diversity. Again, the idea is to develop and standardize a design template for use through a progression of experimental trials. **Figure 1** shows the three-step progression of the trials. **Table 2** shows the results of the first experiment. A nationally representative sample of adult Norwegians was asked if they agree or disagree that "it is important *to acknowledge* the new diversity of Norway." The other half was asked if they agree or disagree that "it is important *to celebrate* the new diversity of Norway" (Ivarsflaten & Sniderman 2017, italics added to highlight contrasting verbs). As **Table 2** shows, a decisive majority of the national majority are willing to acknowledge diversity; far fewer are willing to celebrate it.

The comparative resistance to celebrating diversity points to an upper bound on a willingness to approve of cultural pluralism. For the results of the Ivarsflaten & Sniderman study, though,

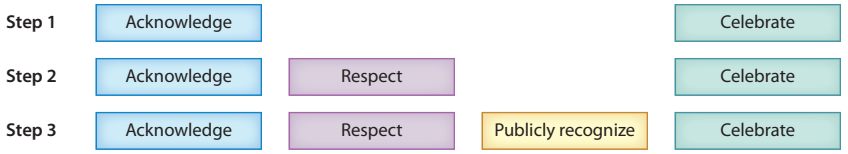

Figure 1

Affirmation of diversity: a sequential factorial design.

the key question is whether majority citizens are ready to go far enough to satisfy the criterion of acknowledgment of worth. What do they mean when they say that it is important to acknowledge the new diversity of Norway? Does it mean no more than that they accept that diversity is now a fact of life, albeit an important one? Or does it connote approval of diversity?

Again, the strategy is to test for synonymy. Operationally, the second experiment in the sequence assesses the interchangeability of "to acknowledge" and "to respect," an unambiguously positive verb. The results, displayed in the middle of **Table 2**, show responses to "to acknowledge" and "to respect" to be interchangeable, and, consistent with the results of the first experiment, both garner more support than "to celebrate." It is worth emphasizing the practice of coarse measurement and again calling attention to the differences in response formats across studies.

Valuing diversity is a necessary but not a sufficient condition to qualify a society as inclusive. Recognition of diversity must also be public (Crowder 2013). The final step in this experimental series is, accordingly, to assess the interchangeability of "acknowledge" or "respect" and "publicly recognize." The third column of **Table 2** reports the results of a study in the United Kingdom. The levels of support for "acknowledge," "respect," and, critically, "publicly recognize" are indistinguishable; for "celebrate," decidedly lower again.

The experimental trials reported in **Table 2** provide a ground-floor argument that national majorities are more open to inclusion than has been recognized. These trials are also an example of a factorial sequence. **Figure 1** offers an illustration of the three-step logic of the sequence. The point to emphasize is that what should be done at step two only becomes obvious when one has seen the results of step one, and what should be done at step three only becomes obvious when one has seen the results of step two. It is also worth remarking that the requirements for replication are progressively stiffened. Step two tests the two parameters of step one; step three the three parameters of step two. All in all, thinking in terms of an experimental sequence is a recommendation for retiring experiment in the singular in favor of experiments in the plural.

Table 2 Acknowledgment of worth of diversity: percentage of respondents in favor of recognizing importance of diversity (Ivarsflaten & Sniderman 2017)

Agree it is important to. . .	Experiment 1: Norway 2[a]	Experiment 2: United States[b]	Experiment 3: United Kingdom 2[c]
Acknowledge diversity	76	73	71
Celebrate diversity	41	61	60
Respect diversity	NA	78	69
Publicly recognize diversity	NA	NA	70

Abbreviation: NA, not asked.

[a]Data source: Norwegian Citizen Panel. Response format: 7-point Likert.

[b]Data source: YouGov. Response format: 4-point Likert.

[c]Data source: YouGov. Response format: 7-point Likert.

VALIDATION REGIMES

There is, if not a crisis, then a justified apprehension about the robustness of results—and, thus, the robustness of the reasoning underpinning them (e.g., Maniadis et al. 2015, Open Sci. Collab. 2015; but see Gilbert et al. 2016). A prime source of the crisis is deception, including self-deception, i.e., believing or persuading yourself that you knew just what result you were after ex ante, even though it only came to mind ex post. Regimes to demonstrate that one has tied one's hands at the start are being put in place, among them preregistration and preanalysis plans. In the next section, I shall explore the contribution that design can make in providing a warrant for results.

Replication requires, minimally, repetition of analysis of the same data or, more ambitiously, repetition of the whole study. Reproducibility entails, minimally, testing the same hypothesis using different measures or, more ambitiously, generating a new prediction to put the same hypothesis to a different test.

Parallel Studies

Consider reproducibility minimally defined, i.e., testing the same hypothesis using different measures. Grimmer and his colleagues (2015) found that a legislator's announcement that she will ask for a benefit for her constituency does her as much good as actually delivering the benefit. Constituents, this suggests, are relatively indifferent to the magnitude of the collective benefits that their representatives secure. If this is so, Grimmer et al. reasoned, constituents will reward legislators as handsomely for bringing home a relatively small amount of money as for securing a full treasure chest. The constituency benefit in this case was a police training program. The range of pay-offs was impressive: The smallest was $10,000, and the largest was $10,000,000, with a uniform distribution drawn between the two. The finding: The size of the grant was associated with increases in favorability rating of legislators' performance up to approximately $1 million; thereafter, a legislator's favorability rating does not vary, however large the amount of money she brings to her constituents.

Two experiments that are grounded in the same reasoning, testing logically equivalent but differentially operationalized predictions, and carried out on independent samples—this is the social science equivalent of walking a high wire without a safety net. The results of the two experiments may differ because the outcome of the first was a fluke. Alternatively, the results may differ because the measures of the second study were inferior to or, perhaps, assessed something different than those of the first. Regrettably, there is no way to choose whether the villain was chance or measurement. Here, then, is a trade-off. The upside of replication is that it can provide gilt-edge evidence that the results of the first trial were a fluke, if indeed they were. The downside of replication is that, beyond confirming that what you thought was true is so, you learn nothing more. The downside of reproducibility is that, if the results of the second experiment are not consistent with those of the first, you cannot tell whether the results were a fluke in the first case or the use of different measures in the second case. The upside of reproducibility is that, if the results of the second experiment are consistent with those of the first, you have made a discovery.

Designing in Pairs

In paired designs, experiments to test the same hypothesis are conducted at the same time. At the same time means on the same survey platform, which in turn means coming up with two different ways to test the same hypothesis twice in the same interview. Reproducibility, not replication, must be the test standard. The catch is that the second experiment is conducted to determine whether the results of the first are robust before knowing the results of the first.

Why pay the cost of designing and conducting two experiments without first determining that the initial experiment is worth following up? The answer turns on the economics of survey research. The principal costs of conducting a public opinion survey—developing an instrument, generating a sample, paying for the machinery to conduct the interview, and lassoing respondents—are incurred before asking the first question. The marginal cost of one more question is trivial. Hence the irresistible urge in public opinion surveys to ask as many questions as possible, consistent with maintaining the quality of responses. The result: most (if not all) of the budget for field research goes to a one-off survey.

Having emptied their pockets on an omnibus survey, researchers must find funds to take the next step. They must complete the product cycle, publishing their results without the ability to resolve questions that now are obvious. Then they must start a whole new product cycle. They must tackle arguably the most challenging mode of scientific writing, a grant proposal; then, assuming fortune smiles, they must collect their data, analyze it, and write up the results of the follow-up study before evidence of the robustness of their initial results stands a chance of seeing the light of day.

A way to avoid this black hole is to design experiments in pairs. An example: policies to assist black Americans are usually justified on racially particularistic grounds—the historical injustices done to black Americans and the discrimination and prejudice that they must still overcome. But the same policies, it was hypothesized, would win more support if justified on morally universalistic grounds (see Skocpol 1992).

Two experiments were designed: one that varied whether the justification for a policy was morally universalistic or racially particularistic; another that varied both the justification and the beneficiary (Sniderman & Carmines 1997). Both experiments were conducted in the same interview. In the first, the same policy, job training programs for black Americans, won more support from individuals on the political left and center when justified in morally universalistic rather than racially particularistic terms. (The justification, universalistic or particularistic, made no difference for those on the political right; nor should it. Just so far as their conservatism was sincere, they ought to oppose a liberal policy.) The pattern of findings of the second experiment was the same as the first. In addition, there was a bonus finding in the second experiment: The same policy that garners the support of only a minority when the beneficiaries or the justification are particularistic wins the support of a majority when both beneficiaries and justification are framed in universalistic terms.

Two points deserve emphasis. First, designing and conducting survey experiments in pairs is as close to costless as one can get, since the interview platform is paid for and a survey experiment characteristically consists of a single question. Second, paired designs hold out the possibility of a big payoff. If successful, researchers need not scrounge funds for a second study to demonstrate that what they showed in the first study is so is so.

Splicing

It is now commonplace for political scientists to argue that replication should be a standard for acceptance of a result. Still, the opportunities to conduct a replication are limited; to publish a replication more limited still. Splicing is a strategy to do the right thing for the right reason—discovery.

Splicing has two parts. The first is to hold constant over experimental trials the template content of the test item. The second is to add new treatments and repeat former ones. Joining together new and old thus allows one simultaneously to extend and to replicate a chain of discovery.

The Suspect Loyalties sequence, shown in **Table 1**, illustrates the dual function of splicing. The initial experiment demonstrated that the political loyalty of Muslim leaders is markedly more

Table 3 The Textbook Experiment: Percentage of respondents in favor of recognizing importance of diversity in textbooks (Ivarsflaten & Sniderman 2017)

Agree textbooks should be written/rewritten	To reflect diversity		To put more emphasis on diversity *(Norway 7 only)*	To put less emphasis on diversity *(Norway 7 only)*
	Experiment 1: Norway 2[a]	*Experiment 2: Norway 7*[a]		
Written	71	69	52	21
Rewritten	55	53	51	20

[a]Data Source: Norwegian Citizen Panel. Response format: 7-point Likert.

likely to be questioned than that of ordinary Muslims. The task, then, was to determine who majority citizens are thinking of when they think of Muslim leaders. Over a series of experimental trials, potential functional equivalents were assayed. The trials went step by step—comparing responses to Muslim religious leaders, Muslim fundamentalists, and Muslim extremists; another compared responses to Muslims, local Muslim politicians, and Muslim female leaders.

In giving an overview of the Suspect Loyalties sequence, I focus on a model of sequential experimentation in which the researcher determines the next step to take from the results of previous steps. Here I want to call out the unique opportunity sequential factorials provide for replication. Repeating the operational template over a sequence of experimental trials opens up an opportunity simultaneously to take a new step and at each step replicate the results of prior steps, administering again some or all of the treatments added incrementally through the sequence of experimental trials. **Figure 1** illustrates how the second step of the Affirmation of Diversity sequence was designed to fully replicate the first, while adding a new experimental condition, and how the third step was designed to fully replicate the first and second steps, while of course adding a new experimental condition. Following this strategy, one can advance and backtrack simultaneously.

The standards for validation can be ratcheted up further. An example is the Textbook experiment progression in the Muslim inclusion study of Ivarsflaten & Sniderman (2017). In the first-stage design, in one experimental condition, respondents are asked whether they agree or disagree that "school textbooks should be *written* to reflect the ethnic and cultural diversity of our country." In the other condition, they are asked whether they agree or disagree that "school textbooks should be *rewritten* to reflect the ethnic and cultural diversity of our country" (italics added to highlight difference between verbs)—a difference of only two letters, "re," but as **Table 3** shows, a difference that makes a difference. On the order of seven in ten Norwegians support textbooks being written to reflect the new diversity of Norway; far fewer, just a bit more than one out of two, support textbooks being rewritten to do exactly the same thing.

The Textbook experiment was repeated in three different countries. In each, the pattern is the same: more support for textbooks being written to reflect diversity; less for their being rewritten to do the very same thing (Ivarsflaten & Sniderman 2017). The results suggest there is more support for inclusion than has been recognized, an optimistic suggestion indeed. It should be a methodological reflex to take the results of public opinion surveys, even if replicated, with a pinch of salt. When the results of the Textbook experiment are set against the undeniable wave of resentment of Muslim immigrants in Western Europe, not to mention the surge of support for nativist parties, a shaker of salt may be in order.

What to do? One answer is to do one's best to show that what one has come to believe is true is in fact false. The challenge is out of the ordinary, admittedly. The starting point is accepting the results of the multiple replications as valid but showing that the conclusion drawn

from them is invalid. What facts, if true, would show that the apparent grounds for optimism are false? Suppose that many who are prepared to acknowledge diversity have mixed feelings. They have come to accept that it is important to acknowledge that people of different backgrounds are part of contemporary life, but they also feel that too much attention is being paid to diversity and not enough to the enduring core of national identity. The Textbook template requires a choice between acknowledging diversity and not acknowledging it. Given only these alternatives, they choose acknowledgment. But given the alternative of acknowledging diversity while placing less emphasis on it than the current fashion does, Ivarsflaten & Sniderman reasoned, that would be their first choice. If so, the optimism the first set of trials suggests should be tempered indeed.

Sequential factorials are not a necessary condition for designing an experiment to show that what one has come to believe is true may be false. But holding constant the operational design of an experiment eliminates a long list of pitfalls. The strategy Ivarsflaten & Sniderman chose was the addition of an option to the Textbook experiment to determine if national majorities, although willing to acknowledge that ethnic and religious diversity is now part of their national narrative, are unenthusiastic about doing so.

The Textbook experiment presents participants with two possibilities, textbooks being written or rewritten. For each possibility, one alternative strictly replicated the first-stage design: "It is important to [write/rewrite] school textbooks to reflect the new diversity of Norway." This time around, though, two other response options were added: to put less emphasis or—for symmetry—to put more emphasis on diversity.

Respondents were assigned at random to one of the six alternatives. The design thus acts as both replication and extension. It is reassuring that, comparing results from the Norway 2 and Norway 7 experiments, one can see that they are, to all intents and purposes, identical. It is the responses to putting "less emphasis" on diversity that point a way forward.

The results are surprising; certainly, they surprised us. We had not imagined that putting more emphasis on diversity would garner more support than putting less emphasis on diversity. But it does. In both arms of the experiment, there is more than twice as much support for putting more emphasis on diversity than less. Indeed, of all the options on offer, putting less emphasis on diversity is far and away the least popular, favored by only a small minority, no more than one out of five. Is it acceptable in a scientific study to report that we were far from thinking that this was a sure thing? The whole point of offering the less-emphasis option was precisely that it stood a good chance of demonstrating that what we had come to believe was true was instead false. But rather than being the most popular option, it was the least.

Putting one's best effort into devising a counter-hypothesis—a conjecture that, if true, demonstrates that what one believes to have established as true is false—is a high-risk strategy. I cannot imagine it being—nor do I believe that it should be—routine. It made sense for this experiment only because it was important to challenge the results of the Textbook experiments as being too optimistic.

THE RESEARCH PROJECT AS A PROCESS

A particular view of science—of how to work through a research question—underpins this review. It is not, I fear, in harmony with the spirit of the times. The current call is for precommitment—public precommitment to hypotheses or least to plans of analysis, formalized ideally before data collection but, in any case, before data analysis. This new regime may prove beneficial, but it is necessary to await evidence. Still, it is interesting that a regulatory regime predicated on bad faith in social science has made such skimpy provision for bad faith in complying with regulation.

It is uncontroversial that research begins in uncertainty. One may be able to give good reasons that a hypothesis is right. But that is not the same thing as being able to demonstrate that it is right. Accomplishing this is the point of research. What has become controversial is the persistence of uncertainty after presentation of results. The fashion now is to conceive of hypothesis testing as a one-off—a decisive demonstration that a claim is or is not valid. Perhaps this is useful. But a different conception of research underpins this review: learning as you go, each advance pointing to a possible new advance. If one can see ahead, one can rarely see far ahead. It is learning what you had not known that allows you to learn (something of) what you still do not know. The research process is a process—a progression of trials. Survey experiments, because of their modesty, lend themselves to this conception of research. Modesty is the operative term: Survey experiments are radically imperfect. Supposedly, no chain of reasoning is stronger than its weakest link. So far as survey experiments are concerned, every link is open to criticism. It is the chain taken as a whole—intuition, inference, measurement guesswork, experimental trial, reconceptualization, more measurement guesswork, and yet another experimental trial—each link helping to hold the link before and after it taut, that is strong.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Alves WM, Rossi PH. 1978. Who should get what? Fairness judgments of the distribution of earnings. *Am. J. Sociol.* 84:541–64

Bansak K, Hainmueller J, Hangartner D. 2016. How economic, humanitarian, and religious concerns shape European attitudes toward asylum seekers. *Science* 354:217–22

Bansak K, Hainmueller J, Hopkins DJ, Yamamoto T. 2017. *Beyond the breaking point? Survey satisficing in conjoint experiments*. Res. Pap. 17-33, Grad. Sch. Bus., Stanford Univ., Stanford, CA

Bansak K, Hainmueller J, Hopkins DJ, Yamamoto T. 2018. The number of choice tasks and survey satisficing in conjoint experiments. *Political Anal.* 26:112–19

Bentzen LE, Bjanesoy L, Ivarsflaten E. 2017. *Patterns of legitimacy on the far right*. DIGSSCORE Work. Pap. Ser., Univ. Bergen, Nor.

Chong D, Druckman JN. 2007. Framing public opinion. *Am. Political Sci. Rev.* 101:637–55

Chong D, Druckman JN. 2010. Dynamic public opinion. *Am. Political Sci. Rev.* 104:663–80

Citrin J, Green DP, Muste C, Wong C. 1997. Public opinion toward immigration reform: the role of economic motivations. *J. Politics* 59:858–81

Crowder G. 2013. *Theories of Multiculturalism*. Cambridge, UK: Polity

Druckman JN. 2001. On the limits of framing effects: Who can frame? *J. Politics* 63:1041–66

Druckman JN, Peterson E, Slothuus R. 2013. How elite partisan polarization affects public opinion formation. *Am. Political Sci. Rev.* 107:57–79

Ford R, Morrell G, Heath A. 2012. "Fewer but better"? Public attitudes to immigration. In *British Social Attitudes: The 29th Report*, ed. A Park, E Clery, J Curtice, M Phillips, D Utting, pp. 26–44. London: Nat. Cent. Soc. Res.

Gamson WA, Modigliani A. 1989. Media discourse and public opinion on nuclear power: a constructionist approach. *Am. J. Sociol.* 95:1–37

Gilbert DT, King G, Pettigrew S, Wilson TD. 2016. Comment on "Estimating the reproducibility of psychological science." *Science* 351:1037

Green PE, Srinivasan V. 1990. Conjoint analysis in marketing: new developments with implications for research and practice. *J. Market.* 54:3–19

Grimmer J, Westwood SJ, Messing S. 2015. *The Impression of Influence*. Princeton, NJ: Princeton Univ. Press
Hainmueller J, Hangartners D, Yamamoto T. 2015. Validating vignette and conjoint survey experiments against real-world behavior. *PNAS* 112:2395–400
Hainmueller J, Hiscox MJ. 2010. Attitudes toward highly skilled and low-skilled immigration: evidence from a survey experiment. *Am. Political Sci. Rev.* 104:61–84
Hainmueller J, Hopkins DJ. 2014. Public attitudes toward immigration. *Annu. Rev. Political Sci.* 17:225–49
Hainmueller J, Hopkins DJ, Yamamoto T. 2014. Causal inference in conjoint analysis: understanding multidimensional choices via stated preference experiments. *Political Anal.* 22:1–30
Ivarsflaten E, Sniderman PM. 2017. *The challenge of inclusion: Muslims in Western Europe*. DIGGSCORE Work. Pap., Univ. Bergen, Nor.
Leeper TJ, Slothuus R. 2016. *If only citizens had a cue: the process of opinion formation over time*. Work. Pap., London Sch. Econ. Political Sci./Aarhus Univ., London/Aarhus, Den.
Maniadis Z, Tufano F, List JA. 2015. How to make experimental economics research more reproducible: lessons from other disciplines and a new proposal. *Replication Exp. Econ.* 18:215–30
Nelson TE, Clawson RA, Oxley ZM. 1997. Media framing of a civil liberties conflict and its effect on tolerance. *Am. Political Sci. Rev.* 91:567–83
Nelson TE, Kinder DR. 1996. Issue frames and group-centrism in American public opinion. *J. Politics* 58:1055–78
Open Sci. Collab. 2015. Estimating the reproducibility of psychological science. *Science* 349(6251):aac4716
Petersen MB, Slothuus R, Stubager R, Togeby L. 2011. Freedom for all? The strength and pliability of political tolerance. *Br. J. Political Sci.* 41:581–97
Rossi P, Sampson W, Bose CE, Jasso G, Passel J. 1974. Measuring household social standing. *Soc. Sci. Res.* 3:169–90
Scheve KF, Slaughter MJ. 2001. Labor market competition and individual preferences over immigration policy. *Rev. Econ. Stat.* 83:133–45
Skocpol T. 1992. *Protecting Soldiers and Mothers: The Political Origins of Social Policy in the United States*. Cambridge, MA: Harvard Univ. Press
Slothuus R. 2010. When can political parties lead public opinion? Evidence from a natural experiment. *Political Commun.* 27:158–77
Sniderman PM, Brody RA, Tetlock PE. 1991. *Reasoning and Choice: Explorations in Political Psychology*. New York: Cambridge Univ. Press
Sniderman PM, Carmines EG. 1997. *Reaching Beyond Race*. Cambridge, MA: Harvard Univ. Press
Sniderman PM, Hagendoorn L. 2007. *When Ways of Life Collide*. Princeton, NJ: Princeton Univ. Press
Sniderman PM, Piazza T. 1993. *Black Pride and Black Prejudice*. Princeton, NJ: Princeton Univ. Press
Sniderman PM, Stiglitz EJ. 2012. *The Reputational Premium: A Theory of Party Identification and Spatial Reasoning*. Princeton, NJ: Princeton Univ. Press
Tomz MR, Weeks JLP. 2013. Public opinion and the democratic peace. *Am. Political Sci. Rev.* 107:849–65
Waddington CH. 1970. *Behind Appearances*. Cambridge, MA: MIT Press

New From Annual Reviews:

Annual Review of Criminology

criminol.annualreviews.org • Volume 1 • January 2018

ONLINE NOW!

Co-Editors: **Joan Petersilia**, *Stanford University* and **Robert J. Sampson**, *Harvard University*

The *Annual Review of Criminology* provides comprehensive reviews of significant developments in the multidisciplinary field of criminology, defined as the study of both the nature of criminal behavior and societal reactions to crime. International in scope, the journal examines variations in crime and punishment across time (e.g., why crime increases or decreases) and among individuals, communities, and societies (e.g., why certain individuals, groups, or nations are more likely than others to have high crime or victimization rates). The societal effects of crime and crime control, and why certain individuals or groups are more likely to be arrested, convicted, and sentenced to prison, will also be covered via topics relating to criminal justice agencies (e.g., police, courts, and corrections) and criminal law.

TABLE OF CONTENTS FOR VOLUME 1:

Annual Review of Political Science

Volume 21, 2018

Contents

Errata

An online log of corrections to *Annual Review of Political Science* articles may be found at http://www.annualreviews.org/errata/polisci