How to enhance the external validity of survey experiments? A discussion on the basis of a research design on political gender stereotypes in Flanders

Paper to be presented at the ECPR General Conference – Prague

Panel 391: Survey Experiments II: From Design to Implementation

Robin Devroe – GASPAR – UGent – robin.devroe@ugent.be

Abstract: (Quasi-) experimental methods are rather scarce in political science, but its use tend to increase in recent years (e.g. Blais, Lachat, Hino, & Doray-Demers, 2011). An experiment is a deliberate test of a causal proposition, typically with random assignment to conditions (Druckman, Green, Kuklinski, & Lupia, 2011b). It is thought that experiments are the best possible way to address the problem of third variables and potentially spurious relationships (Mutz, 2011, p. 9). As the conventional wisdom suggests, experiments are widely valued for their internal validity, but they lack external validity (Mutz, 2011).

External validity is commonly used to refer to the extent to which the causal relationship holds over variations in persons, settings, treatments and outcomes (Shadish & Cook, 2005, p. 83). The concept refers more generally to whether conceptually equivalent relationships can be detected across people, places, times and operationalizations (Anderson & Bushman, 1997, p. 21).

Most political scientists rarely worry about external validity of observational or survey data, but do worry about it with experimental data. The threats to external validity fall into two broad classes: those dealing with generalizations to populations of persons (population validity) and those dealing with the environment of the experiment (ecological validity). In comparison with other methods, experiments do not encounter specific problems regarding population validity. But, when it comes to ecological validity, a weakness of experiments is that they often use artificial stimuli that do not accurately reflect people's everyday world. Since an experimental setting is constructed precisely in order to be internally valid, we cannot be sure that the causal mechanisms hold outside the experiment (Jimenez-Buedo & Miller, 2010; McDermott, 2011).

In this paper, we will focus on the ways in which we can overcome these problems of external validity by discussing the research design of our PhD project on political gender stereotypes in elections. In our experimental design, the central research question is whether political gender stereotypes also prevail in Flanders (Belgium). We will conduct a survey experiment in which respondents will be confronted with hypothetical candidates.

We will discuss how we deal with five different elements to increase external validity in our experimental design: whether the participants resemble the actors who are in real life confronted with these stimuli, whether the context within which actors operate resemble the context of interest, whether the stimulus used in the study resembles the stimulus of interest in the real world, whether the treatments and outcome measures resemble the actual outcomes of theoretical or practical interest and whether the experiment will be replicated across different settings (Druckman, Green, Kuklinski, & Lupia, 2011a).

1. Introduction: experimental research

Druckman et al. (2011c) and Morton and Williams (2008) argue that the experimental study of politics has grown explosively in the past two decades. The drift toward experimentation reflects the value that political and social sciences places on causal inference: experiments facilitate causal inference through the transparency and content of experimental procedures, i.e. the random assignment of observations to control and treatment groups (Druckman, Green, Kuklinski, & Lupia, 2006). In order for one variable to be said to cause another, three conditions generally must be met. This is referred to as the *'holy trinity of causality'*: 1) the two must co-vary, whether over time or across units of analysis; 2) the cause must precede the effect in time; and 3) the relationship between the cause and the effect must not be explainable through some other third variable. The third variable problem is the key reason experiments are known as the gold standard for inferring causality (Mutz, 2011, p. 9).

Druckman et al. define an experiment as *"a study in which the units of observation (typically, subjects, or human participants) are randomly assigned to different treatment or control groups"* (2011c, p. 17). Researchers design experiments to address causal questions. This involves a comparison between a condition in which some sort of intervention is administered (exposing a subject to a stimulus) and another in which this is not the case (i.e. the control condition) (Druckman et al., 2011c).

1.1. Randomization

Experimental research differs from observational research in that the subjects are randomly assigned to different treatments, i.e. potentially causal interventions. Random assignment means that each subject has an equal chance to be in a particular treatment condition (Druckman et al., 2011c). In other words, random assignment ensures that in advance of receiving the treatment, the experimental groups have the same expected outcomes, which is a fundamental requirement for unbiased causal inference (Gerber & Green, 2008). Randomization produces groups that, prior to the experimental intervention, differ with respect to both observable and unobservable attributes only due to chance (Gerber, 2011; Sinclair, 2011) and that are highly similar to one another (Mutz, 2011, p. 138). By comparing the average outcome in the treatment group to the average outcome in the control group,

the experimental researcher estimates the average treatment effect (Druckman et al., 2011c). This allows the researcher to be confident that the only difference across groups or levels is the independent variable. If any changes or differences in the dependent variable are observed, then the researcher can be confident that these are caused by differences in the independent variable (Holbrook, 2011).

In this paper we will focus on the validity of experimental data, and more specifically on their external validity. In the first, theoretical, part of this paper, we will introduce the term validity and elaborate on its different subtypes. Next, we will focus more in depth on the trade-off between internal and external validity. This will be followed by an extensive conceptualization of external validity. We will identify some threats to external validity and we will provide measures to overcome them.

In the second part of this paper, we will apply these theoretical insights to our PhD project on political gender stereotypes. We will first introduce the main principles of our project and elaborate on the experimental design. In the last part of this paper, we will apply the different measures to overcome problems of external validity to our experiment. We end with conclusions.

2. Validity

One of the most important questions for empirical researchers is about the validity of the data: what can we believe about what we learn from the data? A broad definition for 'validity' is provided in the work of Shadish, Cook and Campbell (2002). They use the term validity as the *'approximate truth'* of the inference of a knowledge claim and divide it into four types: statistical conclusion validity, internal validity, construct validity and external validity (Shadish et al., 2002). Statistical conclusion validity is defined as whether there is statistically significant covariance between the variables the researcher is interested in, and whether the relationship is sizeable. Internal validity is defined as the determination of whether the relationship within the particular data-set are causal relationships. Construct validity has to do with how valid the inferences of the data are for the theory (or constructs) that the researcher is evaluating (Morton & Williams, 2008). External validity examines whether or not an observed causal relationship can be generalized to and across different measures, persons, settings and times (Calder, Phillips, & Tybout, 1982).

In practice, however, researchers focus on the difference between internal and external validity. The other subtypes (statistical conclusion and construct validity) are then included under the heading of internal validity. This results in a more narrow definition of validity, which corresponds with the vision of McGraw and Hoekstra (1994). They remark that political scientists have adopted a simplistic view of validity based on the early division of Campbell (1957) using internal validity to refer to how robust

experimental results are within the experimental data, and external validity to refer to how robust experimental results are outside the experiment.

There is a lack of consensus about interrelationships between external validity and the other kinds of validity (Calder et al., 1982). Experiments are for example seen as the best possible design for establishing internal validity, because of their ability to rule out potentially spurious relationships (Mutz, 2011, p. 138), but they lack external validity. A weakness of experiments is that they often use artificial stimuli that do not accurately reflect people's everyday world. Since an experimental setting is constructed precisely in order to be internally valid, we cannot be sure that the causal mechanisms hold outside the experiment (Jimenez-Buedo & Miller, 2010; McDermott, 2011). In this paper, we will focus on this trade-off between internal and external validity and we will try to provide solutions to overcome this trade-off.

2.1. Internal versus external validity

The purpose of an experiment forms the degree to which emphasis should be placed on internal versus external validity. McDermott (2011) argues that internal validity comes first, both sequentially and practically. Without first establishing internal validity, it remains unclear what process should be explored in the real world. External validity follows, as replications across time and populations seek to determine the extent to which these conclusions can generalize (McDermott, 2011).

Campbell (1957) considered an experiment internally valid if the experimenter finds a significant difference between the treatment and control conditions. It refers to the extent to which an experimenter can be confident that his or her findings result from experimental manipulations (McDermott, 2011). The term external validity refers to the approximate validity with which we can infer the presumed causal relationship can be generalized to and across different types of persons, settings and times (Jimenez-Buedo & Miller, 2010).

The conventional wisdom in experimental methods is that there is an inherent tension between internal and external validity (Campbell, 1957; Campbell & Stanley, 1966; McDermott, 2011; Mutz, 2011): an increase in one form of validity may come at the expense of the other. Since an experimental setting is constructed precisely in order to be internally valid, we cannot be sure that the causal mechanisms hold outside the experiment, and therefore, there are grounds to doubt that the phenomena identified under controlled circumstances do hold in the outside world (Jimenez-Buedo & Miller, 2010; McDermott, 2011). The more we ensure that the treatment is isolated from potential confounds, the more unlikely it is that the experimental results can be representative of phenomena of the outside world, since many factors interact (Jimenez-Buedo & Miller, 2010).

McDermott (2011) suggests that there are two principal trade-offs between internal and external validity. First, the balance between these types of validity clearly reflects a difference in value. Concentration on external validity by expanding subject size or representativeness can increase confidence in generalizability, but only to the extent that confounding hypotheses can be excluded from contention. Second, trade-offs between internal and external validity exist in practical terms as well. Enhancing external validity requires more effort because, by definition, the effort must sustain beyond a single study and encompass a sequence of experiments (McDermott, 2011). These trade-offs emerge inevitably over the course of experimental work and depending on the topic, an experimenter may concentrate on maximizing one concern over the other (McDermott, 2011).

It is suggested that internal validity is the more important (Aronson, Carlsmith, & Ellsworth, 1990; Campbell, 1957; Campbell & Stanley, 1966). If random or systematic errors make it impossible for the experimenter to draw any conclusion from the experiment, the question of the generality of these conclusions never arises (McDermott, 2011). McDermott (2011) therefore argues that the concerns regarding external validity are legitimate but they should only arise to the extent that sufficient prior attention has been paid to internal validity.

So far, we have outlined the importance that is accorded to the validity of data and pointed to the inherent trade-off between internal and external validity. Since experiments are widely valued for their internal validity but often criticized for lacking external validity, we have decided to focus on what issues pose a threat to external validity and how this can be overcome. In the next part, we will therefore elaborate on the concept external validity.

2.2. External validity

When Campbell and Stanley (1966) first discussed the concept of external validity, it was clear that the concept involved the generalizability of results beyond the immediate set of conditions observed. External validity refers to the generalizability of findings from a study, or the extent to which conclusions can be applied across different populations, settings, treatments and outcomes (McDermott, 2011; Shadish & Cook, 2005, p. 83). Whether an experimental treatment is generalizable or not refers to whether another manipulation of the same independent variable would produce the same experimental effect (Mutz, 2011, p. 149). It results primarily from replication of particular experiments across diverse populations and different settings, using a variety of methods and measures. The external validity of experimental findings depends upon whether background factors (e.g. subject or setting factors) that are held constant, interact with the manipulated variables (Campbell & Stanley, 1966).

Morton and Williams (2008) consider this as the *'Achilles heel of experiments'*: most political scientists rarely worry about the external validity issue with observational data, but do worry about it with experimental data. There is, however, no deductive or empirical basis for claiming that experimental findings are less generalizable than those of other methods. Just because experiments are high on internal validity does not make them low on external validity (Mutz, 2011). The biggest concern that political scientists focus on with regard to external validity revolves around issues related to the artificiality or triviality of the experimental situation (McDermott, 2011).

To assess the external validity of a causal inference, one must consider from what we are generalizing and to what we aim to generalize. External validity does not simply refer to whether a specific study, if rerun on a different sample, would provide the same results. It refers more generally to whether "*conceptually equivalent*" (Anderson & Bushman, 1997) relationships can be detected across people, places, times and operationalizations (Druckman & Kam, 2011). Mutz (2011) argues that researchers should take into account the following major considerations in evaluating the likely degree of generalizability of an experiment: setting (does the context in which actors operate resemble the real world?), participants (do they resemble the actors who are ordinarily confronted with the stimuli?), measures and treatments (do they resemble the actual outcomes of theoretical or practical interest?). These factors will be discussed more in depth later on. For whatever reason, similarities between research settings and real world settings tend to be emphasized to a much greater degree than the other three factors.

External validity can be examined in multiple ways, including measuring treatment effects in real-world environments, exploring the diverse contexts in which these variables emerge, investigating the various populations it affects and looking at the way basic phenomena might change in response to different situations (McDermott, 2011). Mostly, it is the level of similarity between the real world and the research setting that is held to be key (Mutz, 2011, p. 139). In discussion of internal validity, scholars have used the terms essential and incidental similarity to identify the dimension that are relevant and irrelevant to similarity across experimental conditions. These same terms can be applied to similarities between research settings and their targets of generalization. Some similarities and differences are incidental (irrelevant), while others are essential (relevant) for generalization. It is important to figure out what is essential and what is incidental (Mutz, 2011).

Most concerns that political scientists express regarding external validity reflect their recognition of the artificial nature of the laboratory setting: the trivial tasks presented to subjects offer a poor analogue to the real-world experiences that individuals confront in their daily lives. This characterization of a controlled laboratory experiment reflects a privileging of mundane as opposed to

experimental realism (McDermott, 2011). As Druckman and Kam (2011) argue, experimental realism, as opposed to mundane realism, is critical in assessing the external validity of a research sample.

Experimental realism refers to whether an experiment is realistic, if the situation is involving to the subjects, if they are forced to take it seriously and if it has an impact on them (Druckman & Kam, 2011). Mundane realism refers to the extent to which the experimental setting and procedure resemble things that occur in the real world (Berkowitz & Donnerstein, 1982). One could expect these two kinds of realism to be highly positively correlated, but this is not necessarily the case. It is suggested that experimental realism is of greater importance: low levels of mundane realism constrain the breadth of any generalization, but do not make the study useless (Berkowitz & Donnerstein, 1982; McDermott, 2002, p. 345; Morton & Williams, 2008).

The external validity of a single experiment must be assessed in light of the goal of the study. Theory-oriented experiments are for example not meant to match behaviors observed outside the study, but the key is to generalize to the precise parameters put forth in the given theory. In that way, experiments can still be useful, even if they do not mimic everyday life (Druckman & Kam, 2011).

2.2.1. Threats to external validity

The intent of almost all experimenters is to generalize their findings to some groups of subjects and set of conditions that are not included in the experiment. However, one can identify a number of threats to external validity. The threats to external validity can be called interaction effects, involving X and some other variable (Campbell & Stanley, 1966). In the further development of this paper, we shall discuss several threats to generalizability, and possible solutions for reducing them.

Campbell and Stanley (1966) have pointed to some factors jeopardizing external validity: the interaction effect of testing, in which a pretest might increase or decrease the respondent's sensitivity or responsiveness to the experimental variable; the interaction effect of selection biases, and multiple-treatment inference, likely to occur whenever multiple treatments are applied to the same respondents, because the effects of prior treatments are not usually erasable (Campbell & Stanley, 1966).

The threats to external validity appear to fall into two broad classes: those dealing with generalizations to populations of persons (1) and those dealing with the environment of the experiment (2). The first category refers to the question which population of subjects can be expected to behave in the same way as the experimental subjects, the second to under which conditions, i.e. settings, treatments, experimenters, dependent variables, etc., the same results can be expected. These two broad categories correspond to two types of external validity: population validity (1) and ecological validity

(2) (Bracht & Glass, 1968). Population validity and ecological validity are, however, not independent considerations for designing experiments and interpreting experimental results. Threats to population validity may be the result of some source of ecological invalidity (Bracht & Glass, 1968).

In comparison with other methods, e.g. a classic survey, experiments do not encounter specific problems regarding population validity. One of the purposes of a research study is to learn something about a large group of people by making observations on a relatively much smaller group of subjects. The process of generalizing the experimental results from the sample of subjects to a population is known as statistical inference. Generalization is the ability to make general statements about the effect of some treatment (Bracht & Glass, 1968). Restricted subject population can limit the degree of potential generalizability from all kinds of studies. Selection bias, in terms of nonrandom sampling, is another threat to external validity. If subjects are drawn from a too restrictive or an unrepresentative sample (for example a student sample), then more replication will be required to generalize the results with confidence (McDermott, 2011, p. 36).

In addition to generalizing the results to a population of persons, the researcher wants to be able to say that the same effect will be obtained under other environmental conditions. A generalization assumes that the experimental effect is independent of the experimental environment (which refers to the world ecological: is an effect representative of everyday life?). Here we bump into a problem that applies especially for experimental research methods: one of the weaknesses of experiments is that they often use artificial stimuli that do not accurately reflect people's everyday world. Experiments also typically vary from real world attitude formation and change contexts since information is acquired over a much shorter period of time. Berkowitz and Donnerstein (1982) have, however, argued that external validity does not necessarily require ecological validity. The experimental results may tell something about the conduct of a broad range of people in natural situations even though the subjects and laboratory settings are not physically representative of this population or the real-world situations in which they are embedded.

2.2.2. Measures to enhance external validity

Several strategies can help maximize the potential for increasing relevance and broader applicability of a given experimental study. In general, anything that multiplies the ways in which a particular dynamic is investigated can facilitate prospects for external validity. We focus on the following two strategies: increasing the diversity of circumstances in which a particular phenomenon is investigated and increasing the heterogeneity of the study populations (McDermott, 2011).

In the following part, we will discuss five different topics that have an influence on the external validity of experimental settings: subject selection, context, stimuli, treatment and outcome measures, and

replication. These topics are related to one of the above mentioned strategies: subject selection is related to the heterogeneity of the study population; context, stimuli and replication are related to the diversity of the circumstances.

Subject selection

Random sampling bolsters the external validity of the study insofar as the people in the survey reflect the target population (Druckman et al., 2011c). We have already argued that restricted subject populations can limit the degree of potential generalizability from studies. If subjects are drawn from a too restrictive sample or an unrepresentative sample, then obviously more replication will be required to generalize the results with confidence. When it comes to the representativeness of a sample, some variables may be disregarded, while we must pay attention to some other (eg. age, level of education and gender). It depends on the subject and research questions which variables may be of interest.

The practical and ethical challenges of subject recruitment have led many researchers to rely on convenience samples of college students. Kam et al. (2009) report that from 1990 through 2006, one fourth of experimental articles in general political science journals relied on student subjects, whereas more than seventy percent did so in more specialized journals. For political scientists, who put particular emphasis on generalizability, the use of student participants often constitutes a critical problem for experimental studies (Druckman & Kam, 2011). A crucial question is whether using students as experimental participants creates problems for causal inference. Social scientists in general, and political scientists in particular, view student subjects as a major hindrance to drawing inferences from experimental studies (Druckman & Kam, 2011). According to Druckman and Kam (2011), assessments of how student subjects influence external validity depends on three considerations: 1) the research agenda on which the study builds (e.g. has prior work already established a relationship with student subjects, meaning incorporating other populations may be more pressing?), 2) the relative generalizability of the subjects, compared to the setting, timing and operationalizations and 3) the goal of the study (to build a theory or to generalize one).

Druckman and Kam (2011) argue that there is nothing inherent to the use of student subjects that reduces experimental realism. A student sample only creates a problem when a researcher fails to model a contingent causal effect (i.e. when there is an underlying heterogeneous treatment effect) and when the students differ from the target population with regard to the distribution of the moderating variable. Holbrook (2011) argues, however, that there are important ways in which college undergraduates are different from a generally representative sample. Many political scientists simply assume that a student sample lacks external generalizability (Kam, Wilking, & Zechmeister, 2007).

One of the most important ways to enhance external validity involves increasing the heterogeneity of the study population, McDermott (2011) argues. Including subjects from different age groups, sexes, races and socioeconomic or education statuses, for example, increase the representativeness of the sample and potentiates prospects for generalization. Druckman and Kam (2011) also encourage the use of dual samples of students of nonstudents. The discovery of differences should lead to serious consideration of what drives distinctions. Attention should, however, be paid to the nature of any sample and not just student samples. This includes considerations of nonresponse biases in surveys (Stevens & Ash, 2001).

Context

In experimental designs, the context refers to all background factors that could have an impact on the answers provided by respondents. It is important that the context resembles settings to which the researcher hopes to generalize (Druckman et al., 2011c). Therefore, increasing the diversity of circumstances or situations in which a particular phenomenon is investigated can heighten external validity. Exploring a particular process in a variety of settings can prove particularly helpful for discovering contextual boundaries on particular processes and illustrating the particular dimensions of its operation (McDermott, 2011).

Stimulus

The term stimulus refers to the event or object presented to participants and to which a response is measured. An important issue is whether the stimulus used in the study resembles the stimulus of interest in the real world (Druckman et al., 2011c). Experimental studies on voting behavior should for example resemble a real-world election. King and Matland (2003), in a review of previous experimental work on public evaluations of women candidates, suggest one major limitation of experimental work on this topic: the isolation of candidate gender from other important political and social variables that might influence voter reactions to candidates. If voter evaluations are strongly shaped by a candidate's (or their own) party identification, then there may be no room for heuristics such as gender to ultimately have significant influence (Dolan & Sanbonmatsu, 2011).

Outcome measures and treatments

A fourth important aspect is whether the treatments and outcome measures resemble the actual political outcomes of theoretical or practical interest (Druckman et al., 2011c). A treatment is a set of fixed environmental and institutional conditions (Mutz, 2011), which is administered to the experimental units. Using multiple treatments and measures, or multiple types of measures, as well as

doing everything possible to improve the quality of measures employed, can therefore greatly enhance external validity (McDermott, 2011).

Replication

Morton and Williams (2008) argue that external validity can be achieved if a result can be replicated across a variety of data-sets and situations. They suggest that external validity is about the robustness of the experiments across different formulations (Morton & Williams, 2008). McDermott (2011) also posits that external validity occurs primarily as a function of a strategy of systematic replication. Conducting a series of experiments that include different populations, involve different situations and use multiple measurements establishes the fundamental basis of external validity. A single study, regardless of how many subjects it encompasses or how realistic the environment, cannot alone justify generalization outside the population and domain in which it was conducted (McDermott, 2011).

2.2.3. Conclusion

In sum, external validity covers at least four aspects of experimental design: whether the participants resemble the actors who are ordinarily confronted with these stimuli, whether the context within which actors operate resembles the context of interest, whether the stimulus used in the study resembles the stimulus of interest in the world, and whether the outcome measures resemble the actual outcomes of theoretical or practical interest (Druckman et al., 2011c; Gerber & Green, 2008; Shadish et al., 2002). Following Morton and Williams (2008), we have added a fifth element, i.e. replication: external validity can be achieved if a result can be replicated across a variety of data-sets and situations. The fact that several criteria come into play means that experiments are difficult to grade in terms of external validity, particularly because the external validity of a given study depends on what kinds of generalizations one seeks to make (Druckman et al., 2011c).

3. PhD project on political gender stereotypes

In the second part of this paper, we will focus on the ways in which we can overcome problems of external validity by discussing the research design of our PhD project on political gender stereotypes in Flanders (Belgium). In what follows, we will first introduce the main principles of our PhD project and elaborate on the experimental design. Next, we will apply the different measures to overcome problems of external validity to our experiment.

3.1. Theoretical background and research questions

The starting point of many studies on gender and politics has been the observation that men are overrepresented in politics, whereas women constitute only a small minority of elected officials (Ballington, 2005; Sapiro, 1981; Shvedova, 2005). In our PhD-project, we take the underrepresentation of women in politics as our starting point. It is our aim to search for factors that prevent women from being elected, but our attention will be limited to the role of voters.

More specifically, it is our aim is to highlight the role of **political gender stereotypes** held by voters. Huddy and Terkildsen (1993, p. 120) define political gender stereotyping as "*the gender based ascription of different traits, behavior or political beliefs to male and female politicians*". Voters have different expectations about the issues handled well by male and female politicians, about their character traits and their ideological positions (Alexander & Andersen, 1993; Huddy & Terkildsen, 1993; Matland, 1994; Sapiro, 1981). Following Huddy and Terkildsen (1993), we argue that there are two varieties of political gender stereotypes, those based on women's **traits** and those based on their **beliefs**. According to the trait approach, voters' assumptions about a candidate's gender-linked personality traits drive expectations that women and men have **different areas of issue expertise**. The belief approach suggests that male and female politicians are stereotyped as holding **different political views**.

The existence of political gender stereotypes has been extensively documented in the United States (Alexander & Andersen, 1993; Dolan, 2010, 2014; Fox & Smith, 1998; Huddy & Terkildsen, 1993; Koch, 2000; Rosenwasser, Rogers, Fling, Silvers-Pickens, & Butemeyer, 1987; Sapiro, 1981). Few experimental studies evaluating women candidates have been conducted outside the United States. Since gender differences and gender roles can differ between countries and cultures, we cannot simply transfer these, mostly US-based, findings to other countries and cultures. Therefore, it would be interesting to run a study on the prevalence of political gender stereotypes in continental Europe. We will run this study in Flanders, the Flemish-speaking part of Belgium, which is an interesting case for several reasons and differs from the American political context on several important aspects.

We distinguish between four distinct research questions. First, we would like to find out whether voters translate general stereotyped characteristic traits to the Flemish political sphere; both in terms of issue expertise and beliefs (RQ1). Secondly, we will analyze how the list position of a candidate (a typical feature of proportional representation systems) intervenes with gender cues. The question is whether this additional cue lifts, reinforces or does not affect the perceived general competences of female candidates (RQ2). Third, we will analyze what kind of voters are more prone to hold political gender stereotypes. To that end, we will investigate which individual voter characteristics impact on the existence of gender stereotypes (RQ3). Fourth, we would also like to include the role of issue salience in our design. The prevalence of political gender stereotypes would be less problematic, if the issues for which women are found to be competent are considered as being equally important as the issues for which men are seen as competent. The crucial question in this regard is thus whether issue salience has a positive or a negative impact on the overall evaluation of female politicians (RQ4)?

3.2. Experimental design

3.2.1. Main principles

We will set up an experimental design in which hypothetical candidates are presented to respondents. In brief, these candidates will be presented in text messages in which only their sex, their position on the list and their policy position on a particular issue will be mentioned. We will also mention some arguments to come to these positions, which will always be literally the same for all experimental groups.

Our study will use a 2x3x6 between groups randomized complete block design. We will conduct a repeated measure (within-subjects analysis), i.e. the same respondent will be confronted with different text messages (treatments) (Druckman et al., 2011a). This gives us the advantage that we can collect more data with fewer respondents. Respondents will be randomly assigned to these 6 different treatments. We will alternate the order of the treatments, in order to be able to control for learning or order effects (Chang & Hitchon, 2004). If respondents are able to find out the aim of the experiment (which will be asked at the end of the survey), we will only take the first treatment into consideration.

3.2.2. Online-based survey

We will work with an online-based survey. This is a relatively inexpensive and accessible way to collect data and it brings in the advantage that respondents are not confronted with an interviewer. Therefore we can exclude interviewer effects and do not have to control for this confounding variable.

At the beginning of the online survey, participants will be told that the experiment is a study on political attitudes and that they will be asked to fill out several measures assessing their opinion on various

issues and to evaluate hypothetical political candidates. The hypothetical candidates will only be given a last name. Some set of questions will identify the candidate as “Meneer (mister) Jansssen”, others as “Mevrouw (mistress) Janssen”. In addition, a gender-linked pronoun will be used in the instructional paragraph of the questionnaire. In all other respects, speeches and questionnaires will be identical, in order not to provide any cues to the salience of gender in this study.

There will be different sets of questions. The first set taps the respondents’ evaluations of the candidate’s competence in handling the particular policy areas. Respondents will be asked to rate the candidate’s competence using a fully-labeled 5-point scale. A second set will ask the respondent to indicate what they believe to be the candidate’s ideological position on a fully-labeled 5-point left-right scale. A third set will question the character traits that respondents ascribe to these different candidates. We assume that people assign specific character traits to men and women (women are perceived as being communal, men as being agentic). To make sure that this holds for our set of respondents, we include this as a control question in our survey. We will provide a list of fifteen characteristics and ask respondents to indicate which of them are the five most suitable for the candidate in question. We will also include a set of question to measure whether the respondents would be likely to vote for the candidate.

When it comes to the characteristics of the respondents, the following topics will be included: sex, age, political party identification, ideological position, religion, ethnicity, level of political interest (e.g. frequency of watching the news, reading (online) newspapers), level of education. We will also ask to rate the importance of each policy issue presented on a 10-point scale varying from not at all important to very important. This could provide us a clue about how salient the different policy issues are to the respondents. Finally, the own opinions of the respondents about the issues at stake will serve as a control variable, since this has the potential to influence their evaluation of the candidates.

So far, we have sketched the theoretical background and basic experimental design of our PhD-project on political gender stereotypes. In the next part, we will focus on the external validity of our experimental research project. We will do this by testing how good/bad we score on the different measures that can be taken to enhance the external validity of an experiment. This includes measures regarding the subject selection, the stimulus, the context, the outcome measures and treatments, and replication.

3.3. Testing the external validity of our experiment

3.3.1. Subject selection

Restricted subject population can limit the degree of potential generalizability from studies. In contrast to previous studies, we enhance the external validity of our experiment by conducting the study among

a sample of the population, whereas most other studies analyze students. Although students are of voting age and vary in their level of involvement in politics, they are a more homogeneous population than the general population (Chang & Hitchon, 2004) and they are different from a generally representative sample (Holbrook, 2011). Moreover, it is quite reasonable to hypothesize that students, the youngest voters, may be more liberal in their attitudes toward female candidates (Kahn, 1994). Mutz' (2011, p. 145) view is that there are very few areas of social science in which we can safely assume (without any empirical evidence) that findings based on students are generalizable to other populations.

Working with a population-based survey experiment has therefore obvious advantages in terms of the generalizability of the sample population. We can address generalizability across different segments of the population, far beyond university students, to many educational and socio-economic groups within the population. Including subjects from different age groups, sexes, races and socioeconomic or education statuses increases the representativeness of the sample and potentiates prospects for generalization (Mutz, 2011).

The fact that we will work with an online-based survey provides us the ability to reach diverse populations without (geographical) limitations. Iyengar (2011) argues that, although data make it clear that people who participate in online experiments are not a microcosm of the entire population, the fundamental advantage of online over conventional field experiments cannot be overlooked. Online experiments have the advantage of reaching a participant pool that is more diverse than the pool relied on by conventional experimentalists. Critics have often questioned the extent to which the usual subjects in social science experiments resemble broader, more diverse populations (Mutz, 2011, p. 11). Online techniques permit a more precise targeting of recruitment procedures so as to enhance participants diversity (Iyengar, 2011). We are aware of the fact that working with an online survey entails the risk that we will miss out certain parts of the population, for example the elderly and the less-educated. Therefore, we have to be aware that these groups are equally represented in our final sample of participants. We will make this arrangement with our research facilitator.

Fully operational online opt-in research panels are already available in many European countries, including Belgium (Iyengar, 2011). We have considered the option of randomly selecting respondents from the population register, after obtaining permission from the Privacy Commission. Since this is a costly and rather time-consuming procedure, we think that it might be more useful to request a representative set of respondents with the help of a research facilitator. We will select 1500 respondents spread over the entire Flemish region. We include a number of incentives to obtain a higher response rate: we will provide a participation fee, the online setting makes the threshold to

participate smaller and we will send reminders and follow-up e-mails to the selected respondents. When it comes to this research facilitator, his role will be limited to contacting the respondents and sending follow-up emails. We will prepare our questionnaire in Qualtrics and we will draft the treatments ourselves. In doing so, we try to minimize giving control out of our hands.

A large sample of 1500 respondents gives us the possibility to identify more subtle differences between experimental groups (Mutz, 2011, p. 11). Another advantage is the ability to study specialized subpopulations or to block respondents based on characteristics known to affect the dependent variable (Mutz, 2011, pp. 13-14).

As mentioned before, it is widely acknowledged that systematic differences exist in the ability to contact different kinds of people, as well as their willingness to cooperate. Because non-response and other factors may skew the representativeness of a sample, we can consider the option of weighting. This can help with representativeness when descriptive accuracy is of concern and offers the potential for more accurate estimation of the population average treatment effect (Mutz, 2011, pp. 113-114). Normally, this should not be necessarily in our case, since we have an arrangement with our research facilitator that our sample should be representative on a number of variables (eg. age, gender and level of education). Another option proposed by Mutz (2011, p. 119) is blocking, which means forcing subgroups in the design stage to be equal across control and treatment conditions. This will also be done by the research facilitator.

3.3.2. Context

A second concern is whether the context resembles the settings to which we hope to generalize (Druckman et al., 2011c). At the beginning of our online survey, participants will be told that the experiment is a study on political attitudes and that they will be presented hypothetical political candidates for the next elections of the Flemish Parliament. This is largely in line with the context of real-world elections.

As was suggested by McDermott (2011), exploring a particular process in a variety of settings can be helpful for discovering contextual boundaries on particular processes. We will however not use a variety of settings, since we assume that the prevalence of political gender stereotypes is, among other things, influenced by the institutional context, which can be considered a confounding variable (McGraw, 2011). This refers to the political culture, the electoral system and voting rules. We will hold this constant as the focus is on Flanders' system of proportional representation. We do not wish to extrapolate our results to other regions or to other countries. However, results can be compared with findings from earlier research in other countries afterwards.

Another factor that can be linked to the contextual settings is the duration of experiments (Mutz, 2011). One of the most problematic constraints in doing online experiments is the researcher's inability to know how much time and attention the respondents devote to the treatment. In order to be able to grasp the time devoted to the experiment, text timers will be installed. If respondents stick too long (or too little) on the same page, their responses will not be taken into account since this could mean that they were doing other things in the meantime. The length of the experimental treatment is also important in this regard. We will try to keep the manipulations as short as possible to avoid that respondents lose their attention and to be sure that they read the entire text. This will be tested in our pilot-study. Another way to increase subject engagement is to intersperse the reading material with questions to answer, even if the answers to these question are of little or no use. We will include this kind of questions in the form of manipulation checks, which will be discussed later on.

3.3.3. Stimulus

Another important issue is whether the stimulus used in the experiment resembles the stimulus of interest in the real world (Druckman et al., 2011c). As mentioned before, experimental studies on voting behavior should resemble a real-world election. For presenting the hypothetical candidates, we will be using text messages, which is the standard practice for this kind of studies. Notwithstanding the fact that internet-based surveys open up possibilities for graphics and video as part of the experimental treatment (Mutz, 2011), we have decided to stick to written text messages. Audio- or video-messages may resemble real-world election advertising more, but they are more complicated to set up and the voice and/or physical outlook of the candidates may bias the reaction of the respondents. Next to the text messages, we will provide an image of the ballot, where we indicate the list position of the particular hypothetical candidate. We will do this by putting the dot beside the name of the candidate in red. By doing so, we highlight the list position of the candidate and we enhance the realism of our experiment, since this is also done in election advertising.

As mentioned before, the most important advantage of using experiments is the possibility to control for a number of intervening factors. In order to be able to draw valid conclusions, we have to ensure that only the two variables at stake (gender and list position) play a role in the evaluation made by the respondents. Therefore we will hold the other attributes of the candidates (ethnic origin, age, physical outlook,…) as constant as possible and these will not be discernible in the text messages we will be using. The only variation is constituted by gender and list position, which will be clearly communicated to the respondents. A limitation of our experimental design is, however, that we isolate candidate gender from other important political and social variables that might influence voter reactions to candidates. If voter evaluations are for example strongly shaped by a candidate's (or their own) party

identification, then there is less room for heuristics such as gender to ultimately have significant influence. Sapiro's (1981) study was criticized for not providing the candidate's party affiliation, making the experiment less like a real-world election (Dolan & Sanbonmatsu, 2011). We believe, however, that we offer a methodologically cleaner test (Lammers, Gordijn, & Otten, 2009) by using hypothetical candidates without a partisan affiliation: we are not intervening in actual discussion nor will there be any effect of pre-existing preferences or personal (dis)tastes.

Regarding the contents of the text message, we must take a close look at the policy positions regarding the different issues. These positions will take a central stance (Anckar, 2000). In Belgium's fragmented party system, it is not uncommon that parties take a central stance. Many parties compete for voters' support (Walgrave & De Swert, 2007) and substantive and ideological party differences are smaller in Belgium than in majoritarian countries, e.g. the US. Therefore, respondents will not be surprised by this kind of central/neutral positions. Another advantage is that these positions cannot be easily linked to a particular party, which means that there will be less bias in defining the left-right positioning of the hypothetical candidates.

3.3.4. Outcome measures and treatments

A fourth important aspect is whether the treatments and outcome measures resemble the actual political outcomes of theoretical or practical interest (Druckman et al., 2011c). As suggested by McDermott (2011), using multiple treatments and measures, or multiple types of measures can enhance the external validity of our experimental design.

We will present multiple hypothetical candidates to our respondents. We already mentioned before that our study will use a 2x3x6 between groups randomized complete block design and that we will conduct a repeated measure (within-subjects analysis), i.e. the same respondent will be confronted with different text messages (treatments). The candidate's gender (male versus female) and the list positon (head of list versus position in the middle) are manipulated as treatment variables. We will also include the treatment in which the list position is not mentioned, as a kind of baseline situation, in order to be able to measure the effect of the list position. We will include six different policy issues[1] in our research design. The issues include 2 topics that are, according to previous studies, generally perceived as being communal (e.g. child care and education), topics that are generally perceived as being agentic (e.g. economics and defense) and topics that are generally perceived as being gender-neutral (e.g. the environment and agriculture).

[1] At the moment, we have not decided yet which issues to include, but in order to give an indication, we provide some examples.

In sum, for each policy issue, there will be six experimental groups: a group confronted with a female list puller (1), a male list puller (2), a woman candidate on a middle position (3), a male candidate on a middle position (4), a woman candidate whose list position is not mentioned (5) and a male candidate whose list position is not mentioned (6). This leads us to 36 different treatments and respondents will be randomly assigned to 6 of them. In doing so, we hope to enhance the generalizability of our study through variations in treatments and outcome measures.

We will also try to do everything possible to improve the quality of the treatments and measures employed. We will run pilot tests in order to detect weaknesses in the methodological approach and to remedy them before data gathering takes place. These pilot tests are meant to verify for example whether the different text messages are equally agreeable to read and if their lay-out, word choice and contents are as (gender) neutral as possible.

Another major challenge is creating effective experimental treatments. This is typically measured by a manipulation check (Mutz, 2011). In our online survey, we will also include manipulation checks so that we can verify whether participants were able to correctly answer questions about the sex and the list position, which are the crucial variables at stake, of the candidate whose message they read. These manipulation checks will be included after measuring the dependent variable, so that we do not call undue attention to the stimulus or create a demand for certain kinds of responses.

3.3.5. Replication

A last element regarding the external validity of experiments involves replication. Morton and Williams (2008) and McDermott (2011) suggest that conducting a series of experiments that include different populations, involve different situations and use multiple measurements establishes the fundamental basis of external validity. The central objective of this project is to identify the prevalence of gender stereotypes in Flanders. We have decided to restrict our analysis in a first phase to one monolingual region of Belgium, since it is important that the context in which our research will take place, should be as constant as possible. Although the cultural and institutional differences with Flanders are relatively limited, there are some reasons to expect that the inclusion of Wallonia in our research design makes it less likely that we can control for all variables that determine the (political) context. McDermott (2011) argues that a single study, regardless of how many subjects it encompasses or how realistic the environment, cannot alone justify generalization outside the population and domain in which it was conducted. It is, however, not our aim to generalize our results to other regions or countries.

In a second phase of this project, if we have sufficient time, we would like to run the same experiment in different settings to capture the impact of the political culture and the electoral system. Once our

protocol has been elaborated and applied in Flanders, we could replicate it in other settings, such as Wallonia, or other regions and countries.

4. Conclusion

(Quasi-)experimental methods are increasingly used in political science (e.g. Blais et al., 2011). In this paper, we focused on the generalizability of experimental research results. External validity refers to the extent to which conclusions can be applied across different populations, settings, treatments and outcomes (McDermott, 2011; Shadish & Cook, 2005, p. 83). We have been able to identify a number of threats to external validity, as well as measures to overcome them. The threats to external validity fall into two broad classes: those dealing with generalizations to populations of persons (population validity) and those dealing with the environment of the experiment (ecological validity).

Several strategies can help maximize the potential for increasing relevance and broader applicability of a given experimental study. As suggested by Shadish et al. (2002), external validity covers the following four aspects of experimental design: whether the participants resemble the actors who are ordinarily confronted with these stimuli, whether the context within which actors operate resembles the context of interest, whether the stimulus used in the study resembles the stimulus of interest in the world, and whether the treatments and outcome measures resemble the actual outcomes of theoretical and practical interest. Following Morton and Williams (2008), we have added a fifth element, i.e. replication: external validity can be achieved if a result can be replicated across a variety of data-sets and situations.

We have applied these strategies by testing how good/bad our experimental research project on political gender stereotypes in Flanders scores on the different measures that can be taken to enhance the external validity of an experiment. In contrast to previous studies, we enhance the external validity of our experiment by conducting the study among a sample of the population, whereas most other studies analyze students. Working with a population-based survey experiment has obvious advantages in terms of the generalizability of the sample population. We can address generalizability across different segments of the population, far beyond university students, to many educational and socio-economic groups within the population. Since working with an online survey entails the risk that we will miss out certain parts of the population, e.g. the elderly and the less-educated, we will make sure that these categories are overrepresented in the selection of our sample.

Regarding the context, we are convinced that the setting of our online survey is largely in line with the context of real-world elections. We will however not use a variety of settings, since we assume that the prevalence of political gender stereotypes is, among other things, influenced by the institutional

context, which can be considered as a confounding variable. We will hold this constant as the focus is on Flanders' system of proportional representation.

When it comes to the stimulus, we have to face a trade-off between internal and external validity: the more we ensure that the treatment is isolated from potential confounds, the more unlikely it is that the experimental results can be representative of phenomena of the outside world, since many factors interact. In order to be able to draw valid conclusions, we have to ensure that only the two variables at stake (gender and list position) play a role in the evaluation made by the respondents. Therefore a limitation of our experimental design is that we isolate candidate gender from other important political and social variables that might influence voter reactions to candidates. We believe, however, that we offer a methodologically cleaner test by using hypothetical candidates without a partisan affiliation.

Using multiple treatments and measures can also enhance the external validity of our experimental design. We will present multiple hypothetical candidates to our respondents for six different policy issues. This leads to 36 different treatments which enhances the generalizability of our study. A last element regarding the external validity of experiments involves replication. We have decided to restrict our analysis in a first phase to one monolingual region of Belgium, i.e. Flanders. In a second phase of this project, if we have sufficient time, we would like to run the same experiment in different settings to capture the impact of the political culture and the electoral system. Once our protocol has been elaborated and applied in Flanders, we can replicate it in other settings.

We conclude that our (population-based) survey combines the internal validity of experimental designs with important advantages regarding external validity (Mutz, 2011, pp. 153-154). By simultaneously ensuring internal validity and maximizing the capacity or external validity, this kind of experiments pose an obvious challenge to the basic belief of a trade-off between internal and external validity. By combining representative population samples with rigorous experimental designs, we demonstrate that internal validity can be met without sacrificing the generalizability of the study's sample of participants (Mutz, 2011, p. 131).

5. Bibliography

Alexander, D., & Andersen, K. (1993). Gender as a Factor in the Attribution of Leadership Traits. *Political Research Quarterly, 46*(3), 527-545.

Anckar, C. (2000). Size and party system fragmentation. *Party politics, 6*(3), 305-328.

Anderson, C. A., & Bushman, B. J. (1997). External validity of" trivial" experiments: The case of laboratory aggression. *Review of General Psychology, 1*(1), 19.

Aronson, E., Carlsmith, J. M., & Ellsworth, P. C. (1990). *Methods of research in social psychology*. New York: McGraw-Hill

Ballington, J. (2005). Ten Years of Progress: Enhancing Women's Political Participation. In J. Ballington & A. Karam (Eds.), *Women in parliament: beyond numbers. A revised edition.* (pp. 113). Stockholm: IDEA.

Berkowitz, L., & Donnerstein, E. (1982). External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *American Psychologist, 37*(3), 245.

Blais, A., Lachat, R., Hino, A., & Doray-Demers, P. (2011). The mechanical and psychological effects of electoral systems a quasi-experimental study. *Comparative Political Studies, 44*(12), 1599-1621.

Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. *American educational research journal, 5*(4), 437-474.

Calder, B. J., Phillips, L. W., & Tybout, A. M. (1982). The concept of external validity. *Journal of consumer Research, 9*(3), 240-244.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological bulletin, 54*(4), 297.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Chang, C., & Hitchon, J. C. B. (2004). When does gender count? Further insights into gender schematic processing of female candidates' political advertisements. *Sex roles, 51*(3-4), 197-208.

Dolan. (2010). The impact of gender stereotyped evaluations on support for women candidates. *Political Behavior, 32*(1), 69-88.

Dolan. (2014). Gender Stereotypes, Candidate Evaluations, and Voting for Women Candidates What Really Matters? *Political Research Quarterly, 67*(1), 96-107.

Dolan, & Sanbonmatsu, K. (2011). Candidate gender and experimental political science. In J. N. Druckman, D. P. Green, J. H. Kuklinski, & A. Lupia (Eds.), *Cambridge handbook of experimental political science*. Cambridge: Cambridge University Press.

Druckman, J. N., Green, D. P., Kuklinski, J. H., & Lupia, A. (2006). The growth and development of experimental research in political science. *American Political Science Review, 100*(4), 627-635.

Druckman, J. N., Green, D. P., Kuklinski, J. H., & Lupia, A. (2011a). *Cambridge handbook of experimental political science*. New York: Cambridge University Press.

Druckman, J. N., Green, D. P., Kuklinski, J. H., & Lupia, A. (2011b). Experimentation in Political Science In J. N. Druckman, D. P. Green, J. H. Kuklinski, & A. Lupia (Eds.), *Cambridge handbook of experimental political science*. Cambridge: Cambridge University Press.

Druckman, J. N., Green, D. P., Kuklinski, J. H., & Lupia, A. (2011c). Experiments: an introduction to core concepts. In J. N. Druckman, D. P. Green, J. H. Kuklinski, & A. Lupia (Eds.), *Cambridge handbook of experimental political science*. New York: Cambridge University Press.

Druckman, J. N., & Kam, C. D. (2011). Students as Experimental Participants: A Defense of the'Narrow Data Base'. In J. N. Druckman, D. P. Green, J. H. Kuklinski, & A. Lupia (Eds.), *Cambridge handbook of experimental political science*. Cambridge: Cambridge University Press.

Fox, R. L., & Smith, E. R. (1998). The Role of Candidate Sex in Voter Decision-Making. *Political Psychology, 19*(2), 405-419.

Franzese, R., & Kam, C. (2009). *Modeling and interpreting interactive hypotheses in regression analysis*. Ann Arbor: University of Michigan Press.

Gerber, A. S. (2011). Field Experiments in Political Science. In J. N. Druckman, D. P. Green, J. H. Kuklinski, & A. Lupia (Eds.), *Cambridge handbook of experimental political science*. Cambridge: Cambridge University Press.

Gerber, A. S., & Green, D. P. (2008). Field experiments and natural experiments. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford handbook of political methodology*. Oxford: Oxford University Press

Holbrook, A. L. (2011). Attitude Change Experiments in Political Science. In J. N. Druckman, D. P. Green, J. H. Kuklinski, & A. Lupia (Eds.), *Cambridge handbook of experimental political science* (pp. 141-154). Cambridge Cambridge University Press.

Huddy, L., & Terkildsen, N. (1993). Gender stereotypes and the perception of male and female candidates. *American Journal of Political Science, 37*, 119-147.

Iyengar, S. (2011). Laboratory experiments in political science. In J. N. Druckman, D. P. Green, J. H. Kuklinski, & A. Lupia (Eds.), *Cambridge handbook of experimental political science* (pp. 73-88). Cambridge: Cambridge University Press.

Jimenez-Buedo, M., & Miller, L. M. (2010). Why a trade-off? The relationship between the external and internal validity of experiments. *Theoria. Revista de Teoría, Historia y Fundamentos de la Ciencia, 25*(3), 301-321.

Kahn, K. F. (1994). Does gender make a difference? An experimental examination of sex stereotypes and press patterns in statewide campaigns. *American Journal of Political Science, 38*(1), 162-195.

Kam, C. D., Wilking, J. R., & Zechmeister, E. J. (2007). Beyond the "narrow data base": Another convenience sample for experimental research. *Political Behavior, 29*(4), 415-440.

King, D. C., & Matland, R. E. (2003). Sex and the grand old party an experimental investigation of the effect of candidate sex on support for a republican candidate. *American Politics Research, 31*(6), 595-612.

Koch, J. W. (2000). Do citizens apply gender stereotypes to infer candidates' ideological orientations? *The Journal of Politics, 62*(02), 414-429.

Lammers, J., Gordijn, E. H., & Otten, S. (2009). Iron ladies, men of steel: The effects of gender stereotyping on the perception of male and female candidates are moderated by prototypicality. *European journal of social psychology, 39*(2), 186-195.

Matland, R. E. (1994). Putting Scandinavian equality to the test: An experimental evaluation of gender stereotyping of political candidates in a sample of Norwegian voters. *British Journal of Political Science, 24*(2), 273-292.

McDermott, R. (2002). Experimental methodology in political science. *Political Analysis, 10*(4), 325-342.

McDermott, R. (2011). Internal and external validity. In J. N. Druckman, D. P. Green, J. H. Kuklinski, & A. Lupia (Eds.), *Cambridge handbook of experimental political science* (pp. 27-40). Cambridge: Cambridge University Press.

McGraw. (2011). Candidate impressions and evaluations. In J. N. Druckman, D. P. Green, J. H. Kuklinski, & A. Lupia (Eds.), *Cambridge handbook of experimental political science* (pp. 187). Cambridge: Cambridge University Press.

McGraw, K. M., & Hoekstra, V. (1994). Experimentation in political science: Historical trends and future directions. In M. Delli Carpini, L. Huddy, & R. Y. Shapiro (Eds.), *Research in micropolitics* (Vol. 4, pp. 3-29). Greenwood: JAI Press.

Morton, R. B., & Williams, K. C. (2008). Experimentation in political science. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford handbook of political methodology*. Oxford: Oxford University Press

Mutz, D. C. (2011). *Population-based survey experiments*. Princeton: Princeton University Press.

Rosenwasser, S. M., Rogers, R. R., Fling, S., Silvers-Pickens, K., & Butemeyer, J. (1987). Attitudes toward women and men in politics: Perceived male and female candidate competencies and participant personality characteristics. *Political Psychology, 8*(2), 191-200.

Sapiro, V. (1981). Research frontier essay: When are interests interesting? The problem of political representation of women. *American Political Science Review, 75*(3), 701-716.

Shadish, & Cook, T. (2005). Campbell.(2002). *Experimental and quasi-experimental designs for*.

Shadish, Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton: Mifflin and Company.

Shvedova, N. (2005). Obstacles to women's participation in parliament. In J. Ballington & A. Karam (Eds.), *Women in parliament: Beyond numbers* (pp. 33). Stockholm: IDEA.

Sinclair, B. (2011). Design and analysis of experiments in multilevel populations. In J. N. Druckman, D. P. Green, J. H. Kuklinski, & A. Lupia (Eds.), *Cambridge handbook of experimental political science* (pp. 906). Cambridge: Cambridge University Press.

Stevens, C. D., & Ash, R. A. (2001). The conscientiousness of students in subject pools: Implications for "laboratory" research. *Journal of Research in Personality, 35*(1), 91-97.

Walgrave, S., & De Swert, K. (2007). Where does issue ownership come from? From the party or from the media? Issue-party identifications in Belgium, 1991-2005. *The Harvard International Journal of Press/Politics, 12*(1), 37-67.