AARHUS UNIVERSITY

Coversheet

This is the accepted manuscript (post-print version) of the article.

Contentwise, the post-print version is identical to the final published version, but there may be differences in typography and layout.

How to cite this publication

Please cite the final published version:

Hjortskov, M. (2017). Priming and Context Effects in Citizen Satisfaction Surveys. *Public Administration*, 95(4), 912-926. <https://doi.org/10.1111/padm.12346>

Publication metadata

Title:	Priming and Context Effects in Citizen Satisfaction Surveys
Author(s):	Morten Hjortskov
Journal:	<i>Public Administration</i> , 95(4), 912-926
DOI/Link:	https://doi.org/10.1111/padm.12346
Document version:	Accepted manuscript (post-print)

General Rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

PRIMING AND CONTEXT EFFECTS IN CITIZEN SATISFACTION SURVEYS

Abstract

Citizen satisfaction surveys are used extensively throughout the public sector to assess the performance of public services and to inform decision makers. Recent research points to cognitive biases that may occur in citizens' perceptions of performance of public services, but we know little about possible biases in the collection of these data. This shortcoming is addressed by investigating the priming and context effects that can arise from the structure of citizen surveys – for example from the question-order in the survey. Two independent experimental studies find that prior positively framed questions about police services affect subsequent satisfaction evaluations of other local public services. However, an informational prime about crime and unrelated questions about family-life satisfaction have little effect on the subsequent satisfaction evaluations. The results show that citizen evaluations of public services can be influenced by irrelevant context effects, but this may depend on the topics of prior questions.

INTRODUCTION

The use of citizen satisfaction surveys to measure performance of public services have increased in recent years (Miller *et al.* 2009). This is understandable in light of recent decades' focus on citizens as customers and on performance management (Moynihan 2008). Citizen satisfaction promises an all-encompassing measure that can be used to assess performance in even highly complex services with few accessible objective measures. This makes citizen satisfaction a promising choice for performance measurement. However, recent evidence on cognitive biases in citizen satisfaction and performance evaluations questions the connection between actual performance and citizens' evaluations (Andersen and Hjortskov 2016; Baekgaard and Serritzlew 2016; James and Van Ryzin 2016; Marvel 2016). These insights point to irrelevant influences on performance and satisfaction evaluations which make the results of citizen surveys less promising as performance measures (Stipak 1979). Another potential source of bias arises in the data collection phase due to the way surveys are structured. The influence is sometimes called context effects (Schwarz and Strack 1991). Context effects such as question order and priming of attitudes have been extensively researched in other academic fields and much evidence suggests that they exist (e.g., Tourangeau *et al.* 1989; Higgins 1996).

However, few studies address context effects in citizen satisfaction surveys although they may be a particular challenge in this setting. First, unlike most surveys sent out by researchers, citizen satisfaction surveys are often used as direct performance measures and as feedback to politicians

and decision makers. If context effects bias these particular performance measures, decisions may ultimately be affected as well. Second, because of the ambiguousness of the evaluations citizens are asked to carry out in satisfaction surveys, context effects may potentially have a large influence on the results. Engaging in complex and ambiguous evaluations typically makes people look for ways or cues to make the task easier. The context of the evaluation offers one possibility that may ultimately affect the final result substantially (Schwarz and Strack 1991, p.34; Van de Walle and Van Ryzin 2011, p.1437). Third, citizen surveys are often of an omnibus nature, asking questions about various topics in various orders of appearance. Often, having a survey in the field is an opportunity for many stakeholders in the public administration to obtain knowledge, and as a result, surveys often cover many topics and change over time. Therefore, context effects may be present and widespread in citizen satisfaction surveys and may ultimately have consequences for the decisions made.

This paper studies whether context effects – through prior information and items in a citizen satisfaction survey – affect citizens' evaluations of public services. More specifically, items that are commonly used to measure citizen satisfaction, expectations to and perceived performance of public services, are used as outcome measures in order to assess how prior information and items may affect the results from typical citizen surveys. This is examined in two separate survey experiments carried out in the US through Amazon's Mechanical Turk. The results show that context effects may exist in citizen satisfaction surveys and can substantially affect the results. Specifically, positively framed questions about police services and crime affect subsequent satisfaction questions about public services in a positive direction. This result is replicated in the second study with different respondents. However, an informational prime about crime in the US and questions about family-life satisfaction have no effect on the satisfaction evaluations.

The results add to the ongoing discussion of citizen satisfaction and especially the famed Expectation-Disconfirmation Model (EDM) of formation and antecedents of satisfaction. The EDM assumes that citizens compare their perceptions of the performance of a service with their prior expectations. If the performance falls short of their expectations, they are most likely dissatisfied. If not, they are most likely satisfied (Van Ryzin 2006). Usually, this psychological process is seen as being '*primarily cognitive*' and a result of '*mostly conscious, overt activities*' (Oliver 1993, p.418) and therefore is not subject to unconscious or implicit biases. However, citizens may be under the influence of cognitive biases and sometimes use irrelevant cues or rules-of-thumb when judging and evaluating (Kahneman and Frederick 2005). We know less about when and how biases may occur in citizen satisfaction (Andersen and Hjortskov 2016; Hjortskov 2016), but

context effects in satisfaction surveys may be one source. In contrast to the classic conceptions of the formation of citizen satisfaction in the EDM, cognitive biases may distort the connection between actual performance and satisfaction and contest the use of citizen satisfaction as a performance measure.

The results of the present study indicate that context effects may exist in citizen satisfaction surveys and likely depend on the relevance and type of information citizens are given in the surveys. This adds to the notion that citizen performance and satisfaction evaluations are influenced by cognitive biases. In some of the cases, however, satisfaction and performance evaluations appear quite robust to context effects.

THEORY

Citizen satisfaction surveys are used extensively as measures of performance in the public sector (Miller *et al.* 2009). They are easy to apply, they promise performance measures of services that are not easily captured by objective measures and they represent a ‘citizen-centred’ view of public service performance – all important features in today’s performance management systems (Moynihan 2008). Citizen satisfaction can be conceptualized as evaluative attitudes towards some object or experience (James 2009, p.108), and the citizen satisfaction literature has traditionally focused on the antecedents of satisfaction, such as demographics, trust in government, and service quality (Lyons *et al.* 1992), personal contact to street-level bureaucrats (Hero and Durand 1985), and ideology and local political efficacy (Beck *et al.* 1987).

The EDM, the predominant model of satisfaction formation, states that when prior expectations to a service are compared with the perceived performance of the service, the expectations will either be confirmed or disconfirmed. If they are disconfirmed, the citizen will be disappointed and most likely dissatisfied. If the expectations are confirmed, or exceeded, the citizen will most likely be satisfied (Van Ryzin 2006; James 2009). However, the fundamental assumption of using citizen satisfaction as a performance measure is that perceived performance and satisfaction are closely linked to actual performance of a public service (Stipak 1979; Andersen and Hjortskov 2016).

Context effects in citizen satisfaction surveys have received less attention. It has been known for some time that the structure of surveys influences responses (Tourangeau and Rasinski 1988). For example, respondents can be influenced by numeric response scales (Schwarz, Knäuper, *et al.* 1991), question order (Strack *et al.* 1988; Van de Walle and Van Ryzin 2011; Andersen and

Hjortskov 2016) and whether the preceding items belong to the same topical cluster as the current item (Tourangeau *et al.* 1989; Schwarz, Strack, *et al.* 1991). In other words, attitude measurement can be very context dependent. Just like other attitudes, satisfaction and performance assessments may be affected by structural changes in surveys.

There are two notable exceptions to the lack of interest in these context effects in citizen satisfaction surveys. Van de Walle and Van Ryzin (2011) examine question-order effects in the context of a citizen satisfaction survey. More specifically, they test whether the order of a general satisfaction question and a set of specific satisfaction questions affects evaluations. The conclusion is that overall satisfaction and satisfaction with several specific services is heavily influenced by the question order. Likewise, in a study of priming and question-order effects in a citizen satisfaction survey, Andersen and Hjortskov (2016) find that the question order of the typical overall perceived performance item and the satisfaction item influences the satisfaction evaluation substantively.

In sum, context effects have proven important in general attitude research and may also matter in citizen satisfaction surveys. The next sections explore how context effects work and may come about in citizen satisfaction surveys.

Attitude accessibility, priming and context effects in citizen surveys

Surveys require citizens to solve different tasks. Interpretation, recalling prior opinion and deciding on an answer may obviously be open to many influences, which is described in a large body of literature about survey methodology and the effects of priming (Higgins 1996). Crucial in this process is *attitude accessibility* – the ease with which an attitude comes to mind. The way people view the world is influenced and coloured by needs, desires and attitudes, and an evaluation of real-world objects is necessarily affected, or construed, by the attitudes that are available when the evaluation is taking place (Young and Fazio 2013). In other words, some attitudes are more available than others at different points in time.

Attitude accessibility is thought to have a large influence on how people evaluate specific objects or experiences like a public service experience. For example, a public service like the local police may be perceived (‘construed’) as personal safety or as control and surveillance depending on the accessibility of those construals. The construal of the public service in the specific situation may thus affect evaluations of it.

A specific construal can be activated randomly or through priming via different channels, for example information and prior items in a survey (Higgins 1996; Young and Fazio 2013). This is

thought to produce question-order effects. The sample of beliefs a citizen uses to answer a specific question in a survey will over-represent the beliefs that are most accessible at the specific point in time. Priming through earlier questions in a survey will likely influence this accessibility and in turn affect the answer to the question (Tourangeau *et al.* 1989; Schwarz and Strack 1991; Van de Walle and Van Ryzin 2011).

Three theoretical presumptions are often mentioned for priming or context effects (here called the prime) to carry over to later evaluations (the target). First, the more *ambiguous* an evaluation task is, the more open it will be to context effects (Schwarz and Strack 1991, p. 34). People will possibly look for cues or heuristics that can simplify it. The most accessible attitudes will come to mind first, and accessibility may be influenced by prior information or questions in the survey (Tourangeau *et al.* 1989, p.403). Second, the *topical cluster* of the prime, for example the topic of prior questions in the survey and the target should be related. Priming effects are most pronounced when context and target issues are strongly linked in memory (Tourangeau *et al.* 1989). In citizen surveys about public services, one argument would be that all subjects should be linked in the respondents' memory since they are all about public services. Another would be that most surveys ask for evaluations of diverse services like public education and street cleanliness and citizens are unlikely to view these as belonging to the same topical cluster. It is possible that typical information and prior questions included in citizen surveys can affect later answers to similar as well as less similar topics.

Moreover, the *salience* of the issue may influence the evaluation. An attitude about a highly salient issue may already be accessible, which would render a priming of this issue less effective. In theory, some performance evaluations are more salient than others; say, an evaluation of the police versus an evaluation of public libraries. Priming of a less salient issue may bring out attitudes that would otherwise not be available and cause them to be factored into the next evaluation, which would most likely lead to larger changes in the evaluation. However, attitudes that are not present in the minds of citizens cannot be primed, so one prerequisite of context effects is that citizens have somewhat conflicted beliefs about the issue (Tourangeau *et al.* 1989, p.406).

Finally, one other complexity is worth mentioning. It is not evident whether the *valence* of prior questions affects subsequent evaluations. That is, will it matter if prior questions are framed positively or negatively and is congruence in terms of valence enough for context effects to arise (Fazio 2001, p.119; Klauer and Musch 2003, p.22)? For example, questions in surveys about citizens' evaluations of and attitudes towards the police are sometimes framed in positive ways

like *'I have confidence in the police'* and *'I have great respect for the police'* (Hinds and Murphy 2007, p.42), and sometimes in negative ways like *'police officers stop people for no good reason'* and *'police officers are often rude to the public'* (Schuck and Rosenbaum 2005, p.398). Such framing may ultimately result in different priming effects of prior questions since they may be associated with entirely different experiences and attitudes towards the target. Moreover, a common survey technique to overcome acquiescence effects is to reverse questions in question batteries, which carries the same possibility of priming respondents in a particular direction (Jackman 1973). The fundamental difference between priming through attitude accessibility and framing is basically the difference between *whether* citizens think about an issue and *how* they think about it. The description or presentation of the issue, not the mere mention of it, makes the framing (Scheufele and Tewksbury 2007, p.14).

Context effects in citizen surveys

In sum, priming, both in the form of prior information and prior items in a survey, may affect attitude accessibility and therefore influence answers to later questions in a citizen survey. This should be especially evident when the evaluation is ambiguous, when the target (evaluation of different public services) and the prime (information and prior questions) concern the same topic and when people have conflicted beliefs about an issue. As mentioned, this should be important in citizen satisfaction surveys that often ask respondents to evaluate quite complex and ambiguous services, often consist of evaluations of various topics and are quite diverse in terms of design across organizations and time. Therefore, the expectation is that prior primes and questions will affect subsequent performance and satisfaction assessments and that the framing of prior questions will matter; positively (negatively) framed questions will affect subsequent evaluations in a positive (negative) direction.

The presence of context effects in citizen satisfaction surveys will have theoretical implications for what we know about citizen satisfaction formation and the predominant model of this process: the classic EDM. The EDM fundamentally builds on the assumption that there is a consistent relationship between the citizens' expectations, perceived performance and satisfaction. Moreover, EDM is sometimes described as *'primarily cognitive'* and a result of *'mostly conscious, overt activities'* (Oliver 1993, p.418). This means that citizens are expected to deliberately compare the perceived performance with their prior expectations, and the confirmation/disconfirmation of this effortful process will heavily influence their satisfaction. Therefore, the classic EDM leaves little room for context effects and other biases. This builds upon an idea of *coherence rationality* in the performance and satisfaction evaluations, that is, a rather

strict notion that the system of beliefs and preferences are internally consistent and ‘*immune to effects of framing and context*’ (Kahneman and Frederick 2005, p.277).

However, recent evidence emphasizes the possible biases that may distort the connection between actual performance and citizens’ evaluation of it (Baekgaard and Serritzlew 2016; James and Van Ryzin 2016; Marvel 2016) even when the evaluation is within the paradigm of citizen satisfaction and the EDM (Andersen and Hjortskov 2016). Yet we still know little about when and how cognitive biases affect the EDM. Context effects are another potential source of biases, since the design of citizen surveys should not change citizens’ evaluations of a particular public service; their experiences should be unchanged no matter the order and framing of questions in a survey. If citizen satisfaction surveys, as argued, are prone to context effects in the form of prior primes and questions and the framing of them, it is important to investigate the size and direction of such effects. As far as the possible context effects in citizen satisfaction surveys, we can only look to Van de Walle and Van Ryzin’s (2011) study on the order of specific versus overall questions and Andersen and Hjortskov’s (2016) priming study mentioned above.

STUDY 1

In order to test whether prior questions and information can affect later items in a citizen satisfaction survey, two survey experiments were fielded via Amazon’s Mechanical Turk (MTurk). *Study 1* is a 2X3 factorial experiment with five treatment groups and one control group as illustrated in table 1:

Table 1. *The experimental design of study 1*

	Positive crime and police questions	Negative crime and police questions	No extra questions
Crime prime	Crime prime – positive (1A)	Crime prime – negative (1B)	Crime prime – none (1C)
No prime	No prime – positive (2A)	No prime – negative (2B)	Control (2C)

First, a prime containing information from a recent Gallup poll on Americans’ views on crime (McCarthy 2015) was constructed:

'Now we are going to ask you some questions about the services you receive from your local government'

'One such local service is the local police'

*Crime is an important subject to many Americans. A recent Gallup poll found that **70%** of Americans say there is **more crime in the U.S. now** than there was a year ago, and that **59 %** of Americans describe the **problem of crime** as very or extremely serious' (Gallup, October 2015)*

Half the sample was presented with this prime (the '1' groups). The aim was to increase the accessibility of attitudes concerning crime and police for some respondents. Such information may easily be a part of regular citizen surveys that later ask for the respondents' performance evaluations.

Second, two sets of questions, a positive and a negative version, were asked immediately after the prime (if received). This was done in order to let the respondents engage in attitude rehearsal with both a positively and a negatively valenced version of the questions, which may prime them in two different directions. In both sets, the first part of the questions concerned how the respondents experience safety in their neighbourhoods, the seriousness of crime in their neighbourhoods and their frequency of interaction with the local police during the past six months. The questions are inspired by Weitzer and Tuch's (2005) study on citizen satisfaction with police and Hinds and Murphy's (2007) study on procedural justice and satisfaction with police. The first three questions differed only between the positive and negative treatments in the presentation of the response scales. The positive version presented the positive category first (e.g., very safe, not a problem at all or 0) and the negative version presented the negative category first (e.g., very unsafe, very Serious, 10+).

The next part was a question battery containing five questions about the local police inspired by Schuck and Rosenbaum (2005). The questions and the differences between the positive and the negative framing are illustrated in table 2 and in detail in table B.1 in appendix B. The negative version was taken directly from Schuck and Rosenbaum's study (2005, p.398); for the positive version, the author reformulated these questions to secure as much consistency and as little extra information in the questions as possible. Lastly, a (dis)satisfaction question was added to the four questions in both conditions.

Table 2. *The positive and the negative treatment versions*

Positive Treatment	Negative Treatment
<i>How much do you agree with the following statements?</i>	<i>How much do you agree with the following statements?</i>
Police officers are often respectful to the public	Police officers are often rude to the public
Police officers are verbally polite to people	Police officers are verbally abusive to people
Police officers are physically respectful to people	Police officers are physically abusive to people
Police officers always stop people for good reason	Police officers stop people for no good reason
I am very satisfied with the services provided by the police	I am very dissatisfied with the services provided by the police

Note: Response scale: Strongly disagree, Disagree, Neither Agree nor Disagree, Agree, Strongly agree

In sum, the design contains two types of priming. A typical information prime on the perceptions of crime levels in the US, and two differently framed versions of questions about crime and the police. Both types of priming and the difference between the attitude rehearsals are highly likely to occur in some form in citizen satisfaction surveys used by public organizations. The subsequent outcome questions, which were meant to imitate a traditional citizen satisfaction survey, were taken from Van Ryzin's study on the Expectation-Disconfirmation model (2006) and so was the order of these questions (except the performance battery, which was presented in a randomized order). After a general expectation question, a battery with 11 perceived performance questions was presented. The targets of these evaluations were a broad set of common public services like public education and maintenance of parks and playgrounds. Importantly, two questions on the local police were also included (see table B.2 in appendix B).

The data for study 1 was gathered through an electronic survey fielded via MTurk, an online labour marketplace where respondents for academic research surveys can be recruited. Recently, there has been a surge in research carried out via MTurk probably due to the ease, speed and inexpensiveness of this labour marketplace (Stritch et al. 2017). MTurk respondents (workers) are paid for their efforts and voluntarily sign up for both MTurk and the specific work task, including academic surveys. This makes MTurk workers a quite selective sample. They sign up to earn money and tend to be younger, more liberal, more educated and less employed than the rest of the (US) population which may result in representativeness issues (Paolacci and Chandler 2014). Furthermore, because MTurk workers are paid for their efforts and perhaps mostly are motivated by the payment there have been concerns that some will game the system and click through surveys fast without paying much attention to the questions. This may increase the danger of false positives (Stritch et al. 2017, pp.18–19). However, evidence suggests that MTurk workers are actually quite attentive and estimates obtained through MTurk in some cases are more reliable and valid than other convenience samples like student samples, internet panels and

other traditional survey tools, but perhaps not internet-panels and nationwide probability samples (Berinsky *et al.* 2012; Mullinix *et al.* 2015; Hauser and Schwarz 2016).

A power analysis was carried out on the expectation that effect sizes would be around .25 of a standard deviation. This effect size is typical of survey research in the social sciences because of noise in the measurement (Gerber and Green 2012, p.70). The power analysis conducted in Optimal Design yielded a group size of around 250 individuals (with a power of 80 % and an alpha level of .05). With six groups in the experiment, this amounts to 1500 individuals. The survey for study 1 was pilot tested among 80 workers before the full survey was fielded. Few minor corrections were made in the survey before the full survey was sent to 1,665 workers. Every worker received 0.75\$ for completion of the full survey, which was set to last about 10 minutes (several questions from the survey are not used in this study. See appendix A in the supporting material for an overview of the preceding questions and experiments and tests showing that they have no influence on the main results). The survey was in the field 8-14 December 2015, and 1,608 responses were received (the number of respondents answering the last question in the survey used in the study (overall satisfaction)).

The items in table B.2 are used as overall outcomes (expectations, performance battery, perceived performance, perceived disconfirmation and satisfaction questions). The performance index used in the analysis consists of nine of the performance questions excluding the two police questions (see table B.2). The police questions are left out of the index in order to distinguish these topically close questions from the rest of the index, which concerns evaluations of other services. Results that include them are slightly stronger than results that do not. The items in the index load primarily on the same factor, which has an eigenvalue of 3.5 in a principal component factor analysis and the final index has a Cronbach's alpha value of .80, which is satisfactory. A multinomial logit test of balance with the experimental groups as dependent variable exhibits balance on background variables (not shown) and descriptive statistics on the background and outcome variables are shown in table appendix C, table C.1.

Results: Study 1

In order to check whether the crime prime in study 1 affects key variables, a regression of the crime prime on the citizen evaluations of the personal safety and seriousness of crime in their own neighbourhoods and the number of interactions with the local police is carried out as a manipulation check. Only citizens in the positive/negative treatment groups (1A, 1B, 2A and 2B) answered these questions. The results of this manipulation check presented in table D.1 in appendix D show that there is no effect of the crime prime on these items. The questions used as

dependent variables in this regression are the first questions the citizens in these four conditions encounter after having read the information in the prime, and the topic of the questions should be very close to the information in the prime. In this most likely case, the prime did not appear to affect the citizens and it cannot be expected to affect later evaluation items on other, albeit related, topics.

Table 3 presents the results of the treatments on questions about the performance of the police, the main satisfaction items as they often look in satisfaction surveys (Van Ryzin 2006), and the performance index. These questions pertain to the local services in general while the crime prime and police questions in the treatments focus on a particular service at a general level (not local services in particular). The introduction to the items reads: ‘The next questions are about the local services in general’ and the phrase ‘local government’ is included in all questions. The independent variables are dummies for five of the six treatment groups with the control group (2C) as reference. The performance and satisfaction questions used as dependent variables all appear later in the survey and are scaled 1-4 (models 1 and 2), 1-7 (models 3-6) and an interval performance scale from 1-4 (model 7) respectively (see table B.2).

Table 3. *Effects of treatments on police, satisfaction and performance questions*

	(1) Police Protection	(2) Police- Community	(3) Expectations	(4) Perceived Performance	(5) Perceived Disconfirmation	(6) Satisfaction	(7) Performance Index
Crime Prime + Positive (1A)	0.194** (0.071)	0.270*** (0.073)	0.245* (0.113)	0.247* (0.103)	0.156 (0.095)	0.320** (0.121)	0.137** (0.044)
Crime Prime + Negative (1B)	0.055 (0.070)	0.010 (0.073)	0.163 (0.112)	-0.011 (0.103)	0.067 (0.095)	0.021 (0.120)	0.026 (0.044)
Crime Prime + None (1C)	-0.002 (0.071)	0.039 (0.074)	0.128 (0.113)	-0.077 (0.103)	0.044 (0.096)	-0.035 (0.121)	0.030 (0.044)
No Prime + Positive (2A)	0.165* (0.070)	0.168* (0.073)	0.144 (0.113)	0.192 (0.103)	0.163 (0.095)	0.231 (0.121)	0.110* (0.044)
No Prime + Negative (2B)	0.010 (0.071)	0.071 (0.074)	-0.008 (0.113)	-0.012 (0.103)	0.076 (0.095)	0.011 (0.121)	-0.020 (0.044)
Constant	2.661*** (0.050)	2.463*** (0.053)	4.292*** (0.081)	4.560*** (0.074)	4.136*** (0.068)	4.739*** (0.086)	2.658*** (0.031)
<i>N</i>	1608	1608	1608	1608	1608	1608	1608

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. OLS models with the control group (2C) as reference. The outcome of models 1 and 2 are scaled from 1-4 and outcomes of models 3-6 are scaled 1-7. The performance index (model 7) is scaled from 1-4 (interval) and consist of nine performance items (excluding the two police items from models 1 and 2). Standard errors in parentheses.

Table 4. *Effects of combined treatments (factors)*

	(1) Police Protection	(2) Police- Community	(3) Expectations	(4) Perceived Performance	(5) Perceived Disconfirmation	(6) Satisfaction	(7) Performance Index
Crime Prime (1A, 1B, 1C)	0.024 (0.040)	0.027 (0.042)	0.134* (0.064)	-0.007 (0.059)	0.009 (0.054)	0.022 (0.069)	0.035 (0.025)
Positive Police (1A, 2A)	0.181*** (0.050)	0.200*** (0.052)	0.131 (0.079)	0.259*** (0.072)	0.137* (0.067)	0.294*** (0.085)	0.108*** (0.031)
Negative Police (1B, 2B)	0.034 (0.050)	0.020 (0.052)	0.014 (0.079)	0.028 (0.072)	0.050 (0.067)	0.034 (0.085)	-0.012 (0.031)
Constant	2.648*** (0.041)	2.469*** (0.043)	4.289*** (0.065)	4.524*** (0.060)	4.154*** (0.055)	4.710*** (0.070)	2.656*** (0.025)
<i>N</i>	1608	1608	1608	1608	1608	1608	1608

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. OLS models with the control group (2C) as reference. The outcome of models 1 and 2 are scaled from 1-4 and outcomes of models 3-6 are scaled 1-7. The performance index (model 7) is scaled from 1-4 (interval) and consist of nine performance items (excluding the two police items from models 1 and 2). Standard errors in parentheses.

Model 1 and 2 test whether questions about subjects closely linked to the information and prior questions are affected by the treatments. Two different performance questions were asked with direct relevance for the crime and police treatments: One about police protection and one about police-community relations. Both are strongly affected by the positively framed crime and police questions (1A and 2A). The performance of police protection increases by .194 on the 1-4 scale when both the crime prime and the positive police questions are presented (1A) and by .165 if only the positive police questions are asked (2A). In the case of the police-community question, the crime prime with the positive police treatment results in a significant effect of .270, whereas the positive police treatment alone results in an effect of .168. The other treatment groups do not have significant effects on these police questions.

It seems that the crime prime has little influence and that it is the positive police treatments that drive the results, which is also expected given the above result from the manipulation check. Table 4 analyses the three different factors in the 2X3 factorial experiment on the same outcomes as in table 3 by using indicators for being in the groups with the crime prime treatment (1A, 1B and 1C), the positive police treatment (1A and 2A) or the negative police treatments (1B and 2B). It is evident from table 4 that neither the crime prime nor the negative police treatment had an effect on the perceived performance of police protection and police-community relations (models 1 and 2). Likewise, there is no significant difference between the positively and the negatively framed police questions (not shown). However, the positive police questions had a substantial effect on the answers to both questions in a positive direction with an effect of .181 on the police protection questions and .200 on the police-community question.

Looking at the satisfaction questions in model 3-6 in table 3, only the treatment group with both the crime prime and the positive police questions (1A) affects these outcomes significantly. This is true for citizen expectations, perceived performance and satisfaction (models 3, 4 and 6), but not perceived disconfirmation (model 5). None of the other treatment groups seem to have an independent effect. However, when the treatment groups are combined into factors in table 4, the positive police factor turns out significant in the cases of perceived performance, perceived disconfirmation and satisfaction with estimates of .259, .137 and .294 on the 1-7 scale respectively. The positive police questions have no effect on citizen expectations. However, the crime prime increases citizen expectations by .134, which is a significant effect. This may reflect the gain in power with this approach since the coefficients are almost the same (.134 versus .128) as when the prime-only condition (1C) is compared with the control group (2C) in table 3.

The same pattern emerges for the performance index (formed without the two police performance questions, model 7). The condition with the positive police questions (1A) has a significant positive effect on the performance index in table 3. The coefficient is .137 on the interval scale from 1 to 4. The other conditions in the experiment do not seem to affect the performance evaluations of specific local services. The combined effect of the crime prime in table 4 is not significant either, but the combined positive police treatment is significant at the 1 % level with a coefficient of .108.

Figure 1 further explores what drives the effects in the performance index by regressing each performance item on the different treatment groups. The results on the performance index (table 3, model 7) are included for reference as are the two performance questions on the police that are not included in the performance index (table 3, models 1 and 2). The results are shown as coefficient plots in which estimates are illustrated by circles, and their confidence intervals (5%) are illustrated by spikes. Significant results are illustrated by circles and spikes that do not touch the dotted line at 0 on the scales at the bottom of the figure.

Figure 1. *Effect of treatments on individual performance questions*

Note: OLS regression with the control group (2C) as reference. Spikes represent the 5 % confidence interval. Circles represent the point estimates. Items are scaled from 1 to 4 (ordinal scale for single items, interval scale for the performance index). N = 1,608.

In many cases, the pattern from the performance index is evident in the single items in figure 1. Especially the treatment group that received both the prime and the positively framed police questions (1A) has different performance evaluations compared with the control group. This is also true for the performance evaluations of street maintenance, parks and playgrounds, public libraries, fire protection and public education. The treatment that only contains the positively framed police questions (2A) has significant positive effects on the performance evaluations on parks and playgrounds and public education at the 5 % level.

STUDY 2

Study 1 shows that some (positive) types of questions asked early on in a citizen satisfaction survey may affect subsequent evaluations. Positively framed crime and police questions affected later citizen satisfaction and performance evaluations of local public services. Even services that are not directly related to crime and police were affected. The reason may be that the police deliver a public service and the evaluation therefore is carried out within the same topical cluster. Another reason may be that the police are a salient topic and it is framed positively. *Study 2* seeks to replicate the results from the positive police question without a crime prime in study 1 (2A). Furthermore, it varies on one part of the treatment in the positive police case, which may explain the effect – the common topic of public services between the prime and the target – while seeking to hold the positive frame and high saliency constant. Study 2 incorporates five questions from Zabriskie and Ward’s satisfaction with family-life scale (2013) to accomplish this. The questions can be seen in table 5.

Table 5. *The family-life satisfaction scale*

Family-life satisfaction scale
In most ways my family life is close to ideal
The conditions of my family life are ideal
I am satisfied with my family life
So far I have gotten the important things I want in my family life
If I could live my family life over, I would change almost nothing
Note: Response scale: Strongly disagree, Disagree, Neither Agree nor Disagree, Agree, Strongly agree

This is to test for context effects from a topic that is not a public service but presumably has high salience and is positively framed just like the positive police questions. The basic design is illustrated in table 6:

Table 6. *The experimental design of study 2*

Treatment	Positive crime and police questions	Satisfaction with family-life scale	No extra questions
Name	Positive police	Family	Control

Study 2 was also carried out via an MTurk survey among US citizens. After a pilot test among 70 workers, the survey was sent to 1,857 workers without changes. Every worker received 0.80\$ for completing the survey, which was set to last about 12 minutes (again, several questions from both surveys are left out in this study, see appendix A). The survey containing study 2 was in the field 18-29 September 2016 and received 1,186 responses (the number of respondents answering the last question in the studies in the survey (overall satisfaction)). Due to a technical error, one treatment group is incomparable to the control group and is left out of the analysis). Again, a multinomial logit test of balance returns acceptable results (not shown), and descriptive statistics on background and outcome variables are shown in table C.2 in appendix C. The performance index has good reliability (Cronbach's alpha = .79) and loads primarily on a single factor in a principal component factor analysis (eigenvalue = 3.4).

Results: Study 2

Table 6 shows the effect of the two treatments' positive police and family-life satisfaction questions on the performance and satisfaction questions. The positive police questions have a positive effect on the police protection, police-community, expectations, perceived performance and satisfaction questions. Likewise, there is a significant effect of the positive police treatment on the performance index with an estimate of .092, which is quite close to the estimate of .110 in table 3 (model 7). All except perceived disconfirmation are significant at the 5 % level. These results to a large degree replicate the results from study 1. The only significant effect of the family treatment is on perceived disconfirmation. The estimate is .149 and it is significant at the 5 % level.

Table 6. *Effect of treatments on police, satisfaction and performance questions*

	(1) Police Protection	(2) Police- Community	(3) Expectations	(4) Perceived Performance	(5) Perceived Disconfirmation	(6) Satisfaction	(7) Performance Index
Positive Police	0.170** (0.058)	0.190** (0.061)	0.273** (0.100)	0.180* (0.086)	0.059 (0.075)	0.246* (0.100)	0.092* (0.037)
Family-life Satisfaction	0.063 (0.057)	0.083 (0.061)	0.084 (0.099)	0.085 (0.085)	0.149* (0.074)	0.104 (0.099)	0.058 (0.036)
Constant	2.656*** (0.040)	2.485*** (0.043)	4.054*** (0.070)	4.488*** (0.060)	4.144*** (0.052)	4.619*** (0.070)	2.615*** (0.026)
<i>N</i>	1187	1187	1187	1187	1186	1186	1187

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. OLS models with the control group as reference. The outcome of models 1 and 2 are scaled from 1-4 and outcomes of models 3-6 are scaled 1-7. The performance index (model 7) is scaled from 1-4 (interval) and consist of nine performance items (excluding the two police items from models 1 and 2). Standard errors in parentheses.

Study 2 shows that it is possible to replicate the main finding from study 1 and that this finding may be the result of the topical cluster of both the treatment and the outcome questions— public services – being related rather than of the salience of the police and crime topic. At least family-life satisfaction, which is clearly not a public service but was framed positively and should be highly salient like the positive police treatment, does not influence later evaluations of public services.

DISCUSSION AND CONCLUSION

This paper has argued that context effects in surveys may arise in evaluation tasks when an evaluation is ambiguous, when the topics of the context questions and information are closely related to the evaluation target's topic and when respondents have somewhat conflicted beliefs about the issue that can be primed. Citizens might be subject to context effects when they answer citizen satisfaction surveys, because these are often of an omnibus nature and change over time and because they ask citizens to evaluate ambiguous services. The central motivation behind this study is that context effects in citizen satisfaction surveys would be problematic since results from such surveys are often used as performance information that potentially influences decision makers and politicians and their actual decisions.

The contribution of the study is that it investigates context effects in citizen satisfaction surveys and takes their typical structure, topics and length into account in its experimental variations of prior questions, information and framings. The study found significant context effects on citizen expectations, performance evaluations of different public services and satisfaction in a typical citizen satisfaction survey given to two samples of MTurk workers. This is found among citizens who have actual experiences with the services they evaluate. Priming respondents with a set of five positively framed questions about the police affected many of the subsequent evaluations although some of them were placed much later in the survey and arguably concerned qualitatively different evaluations. However, not all treatments affected the evaluations. Negatively framed police questions and a prime about perceptions of crime in the US had limited effect and hence citizen surveys seem robust to these particular kinds of context effects. The effects of the positive police questions are replicated in study 2, which underlines the relevance of this result. Study 2 also finds that positively framed questions about family-life satisfaction generally do not affect later citizen evaluations.

Before the results and their implications are discussed further, a few caveats should be mentioned. First, the effects of the positively framed police questions are relatively small.

Significant effects are in the range of .18-.32 for the 7-point items (overall expectations, perceived performance, perceived disconfirmation and satisfaction) and .09-.14 for the performance index. The different estimates have a Cohen's *d* of approximately .2, which is sometimes labelled as small effects (Cohen 1988). However, large effects (according to the Cohen framework) are rare in the social sciences (Murnane and Willett 2011, pp.91–92), and these effects are evident across multiple questions placed far from each other in the survey, even though priming effects are known to wear off quickly (Tourangeau *et al.* 1989). Also, due to possible error in measurement, standard errors increase with smaller effect sizes as a consequence. This is a typical challenge in measuring attitudes in surveys (Gerber and Green 2012, p.70). Second, although an advantage for the realism, a design that seeks to imitate the setup in typical citizen satisfaction surveys is, at least in this case, not fit for a more rigorous test of the mechanisms in the context effects. Third, the use of MTurk workers, who may be motivated by the payment more than the topics of the questions, is still partly controversial (Stritch *et al.* 2017). Therefore, research should seek to replicate the findings with other samples of citizens as well.

The results of this study have implications for how citizen surveys should be structured. When the topic of prior questions in the survey is related to later evaluations, context effects may appear and influence questions that are later used for performance feedback to decision-makers. In the present case, the relation is presumably established because all evaluations and context questions concern public services. Conversely, the study also finds that prior questions evaluating an entirely different subject, which is not a public service, does not influence later evaluations of public services even though these questions were also about a salient topic and positively framed. This underlines the theoretical assertion that a close relation between context and target is fundamental to context effects.

Framing of prior questions did not seem to have a differential effect on later evaluations. Only the positive police questions affected later evaluations significantly, while the negatively framed did not (there was no significant difference between the two). In other words, there is no sign of a negativity bias here (Soroka 2014) with larger effects of the negative treatment. Rather, there are signs of a 'positivity bias'. Therefore, this study does not add much to the debate about the direction and valence of context effects (e.g., Tourangeau and Rasinski 1988; Fazio 2001). This may be due to the average prior attitudes of the respondents. If citizens in general are supportive of the police, the positively framed police questions may make these positive attitudes much more accessible to them, and accessibility of positive attitudes about public services will affect

answers to later questions about public services. If negative attitudes about the police are largely absent to begin with, the prime will have little effect according to the theory.

However, this argument is complicated by the crime questions that are also part of the treatment. The two topics are highly related and very likely to be presented together in a citizen satisfaction survey, but they also speak in different directions. Citizens will likely have negative attitudes towards crime, and following the above argument the negative treatment should have the most influence. What we observe from the experiments is the average effect, and the police questions, which are presented last in the treatment, may just have the most influence, and therefore we only observe an effect from the positive police questions. Moreover, positively framed family-life satisfaction questions did not have an effect on the subsequent evaluations, so positivity per se does not seem to do the trick either. Further research is needed to determine these framing mechanisms in depth to supply practitioners with actionable advice on the structure of surveys in this regard.

Moreover, it can be argued that police services and crime are high saliency topics for many citizens and cause stronger reactions than other topics. Higher associative strength would influence attitude accessibility and therefore the likelihood that the attitude can be activated automatically (Fazio 2001, p.122). However, high saliency does not necessarily make for a stronger prime, since such attitudes may already be accessible. Citizens may be pre-treated on these subjects, whereas a prime about a low-saliency subject would bring attitudes to the table that would not otherwise be taken into account. There is no effect of the negatively framed police questions, however, and they presumably have the same saliency as the positive ones. Another supposedly high saliency treatment, the family-life satisfaction questions, also did not affect the satisfaction variables (apart from perceived disconfirmation). Further research should test whether different topics with different salience have different context effects on citizens' evaluations of public services (for example, police versus library services).

The literature has limited practical suggestions for designing citizen surveys to avoid context effects. However, a few could be mentioned. First, randomization of the question order whenever possible and meaningful should offset context effects that pertain to the order of questions. This is often doable given the now common practise of conducting online surveys. Second, in a few cases it has been shown that prior questions or information may debias later questions (Tourangeau *et al.* 1989, p.405). For example, a question about the weather can strip a later question on life satisfaction of the influence of the current weather and mood (Schwarz and Clore 1983). Likewise, open-ended questions may make respondents think more carefully about a

subject before answering the closed questions. Using open-ended questions in this way has also been shown to remove the bias in some cases (Schwarz and Clore 1983; Andersen and Hjortskov 2016).

In general, motivating respondents to think carefully about the subject and evaluation at hand seems to be the key to avoiding irrelevant influences in evaluations, whether it is done by priming their ‘desire for accuracy’ (Schuette and Fazio 1995), priming them to consider information in an even-handed manner (Bolsen *et al.* 2014), or holding them accountable by warning them that they will be asked to justify their evaluation later in the survey (Tetlock 1983). The difficult thing here is to know when and how context effects might affect performance and satisfaction measures and then try to design surveys to avoid it without inducing even more context effects.

LITERATURE

- Andersen, S.C., Hjortskov, M. (2016) ‘Cognitive Biases in Performance Evaluations’, *Journal of Public Administration Research and Theory*, 26(4), 647–662.
- Baekgaard, M., Serritzlew, S. (2016) ‘Interpreting Performance Information: Motivated Reasoning or Unbiased Comprehension’, *Public Administration Review*, 76(1), 73–82.
- Beck, P.A., Rainey, H.G., Nicholls, K., Traut, C. (1987) ‘Citizen views of taxes and services: A tale of three cities’, *Social Science Quarterly*, 68(2), 223.
- Berinsky, A.J., Huber, G.A., Lenz, G.S. (2012) ‘Evaluating online labor markets for experimental research: Amazon. com’s Mechanical Turk’, *Political Analysis*, 20(3), 351–368.
- Bolsen, T., Druckman, J.N., Cook, F.L. (2014) ‘The Influence of Partisan Motivated Reasoning on Public Opinion’, *Political Behavior*, 36(2), 235–262.
- Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*. 2nd Edn. Hillsdale, New Jersey: Laurence Erlbaum Associates.
- Fazio, R.H. (2001) ‘On the automatic activation of associated evaluations: An overview’, *Cognition & Emotion*, 15(2), 115–141.
- Gerber, A.S., Green, D.P. (2012) *Field Experiments: Design, Analysis, and Interpretation*, W. W. Norton: New York.
- Hauser, D.J., Schwarz, N. (2016) ‘Attentive ‘Turkers’: MTurk participants perform better on online attention checks than do subject pool participants’, *Behavior Research Methods*, 48(1), 400–407.
- Hero, R.E., Durand, R. (1985) ‘Explaining Citizen Evaluations of Urban Services A Comparison of Some Alternative Models’, *Urban Affairs Review*, 20(3), 344–354.
- Higgins, E.T. (1996) ‘Knowledge activation: Accessibility, applicability, and salience.’, in Higgins, E.T. and Kruglanski, A.W., eds., *Social Psychology. Handbook of Basic Principles*, Guilford Press: New York, NY, 133–168.
- Hinds, L., Murphy, K. (2007) ‘Public Satisfaction With Police: Using Procedural Justice to Improve Police Legitimacy’, *Australian & New Zealand Journal of Criminology*, 40(1), 27–42.
- Hjortskov, M. (2016) *Citizen Satisfaction: Political Voice and Cognitive Biases*, Aarhus, Forlaget Politica. Available:<http://politica.dk/forlaget/politicas-phd-serie/morten-hjortskov-larsen/> [accessed 2 May 2017].
- Jackman, M.R. (1973) ‘Education and Prejudice or Education and Response-Set?’, *American Sociological Review*, 38(3), 327–339.

- James, O. (2009) 'Evaluating the Expectations Disconfirmation and Expectations Anchoring Approaches to Citizen Satisfaction with Local Public Services', *Journal of Public Administration Research and Theory*, 19(1), 107.
- James, O., Van Ryzin, G.G. (2016) 'Motivated Reasoning about Public Performance: An Experimental Study of How Citizens Judge the Affordable Care Act', *Journal of Public Administration Research and Theory*, muw049.
- Kahneman, D., Frederick, S. (2005) 'A Model of Heuristic Judgment', in Holyoak, K.J. and Morrison, R.G., eds., *The Cambridge Handbook of Thinking and Reasoning*, Cambridge University Press: Cambridge, New York, 267–294.
- Klauer, K.C., Musch, J. (2003) 'Affective priming: Findings and theories', in Musch, J. and Klauer, K.C., eds., *The Psychology of Evaluation: Affective Processes in Cognition and Emotion*, 7–49.
- Lyons, W.E., Lowery, D., DeHoog, R.H. (1992) *The Politics of Dissatisfaction: Citizens, Services and Urban Institutions*, Armonk, NY: M. E. Sharpe.
- Marvel, J.D. (2016) 'Unconscious Bias in Citizens' Evaluations of Public Sector Performance', *Journal of Public Administration Research and Theory*, 26(1), 143–158.
- McCarthy, J. (2015) More Americans Say Crime Is Rising in U.S. [online], *Gallup.com*, available: <http://www.gallup.com/poll/186308/americans-say-crime-rising.aspx> [accessed 26 Nov 2015].
- Miller, T.I., Kobayashi, M.M., Hayden, S.E. (2009) *Citizen Surveys for Local Government*, 3rd ed., International City Management Association.
- Moynihan, D.P. (2008) *The Dynamics of Performance Management: Constructing Information and Reform*, Georgetown University Press.
- Mullinix, K.J., Leeper, T.J., Druckman, J.N., Freese, J. (2015) 'The Generalizability of Survey Experiments', *Journal of Experimental Political Science*, 2(02), 109–138.
- Murnane, R., Willett, J. (2011) *Methods Matter: Improving Causal Inference in Educational and Social Science Research*, OUP USA: Oxford ; New York.
- Oliver, R.L. (1993) 'Cognitive, affective, and attribute bases of the satisfaction response', *Journal of consumer Research*, 418–430.
- Paolacci, G., Chandler, J. (2014) 'Inside the turk understanding mechanical turk as a participant pool', *Current Directions in Psychological Science*, 23(3), 184–188.
- Scheufele, D.A., Tewksbury, D. (2007) 'Framing, Agenda Setting, and Priming: The Evolution of Three Media Effects Models', *Journal of Communication*, 57(1), 9–20.
- Schuck, A.M., Rosenbaum, D.P. (2005) 'Global and Neighborhood Attitudes Toward the Police: Differentiation by Race, Ethnicity and Type of Contact', *Journal of Quantitative Criminology*, 21(4), 391–418.
- Schuetz, R.A., Fazio, R.H. (1995) 'Attitude accessibility and motivation as determinants of biased processing: A test of the MODE model', *Personality and Social Psychology Bulletin*, 21(7), 704–710.
- Schwarz, N., Clore, G.L. (1983) 'Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states.', *Journal of personality and social psychology*, 45(3), 513.
- Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., Clark, L. (1991) 'Rating scales numeric values may change the meaning of scale labels', *Public Opinion Quarterly*, 55(4), 570–582.
- Schwarz, N., Strack, F. (1991) 'Context effects in attitude surveys: Applying cognitive theory to social research', *European review of social psychology*, 2(1), 31–50.
- Schwarz, N., Strack, F., Mai, H.-P. (1991) 'Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis', *Public opinion quarterly*, 55(1), 3–23.
- Soroka, S.N. (2014) *Negativity in Democratic Politics: Causes and Consequences*, Cambridge University Press.

- Stipak, B. (1979) 'Citizen satisfaction with urban services: Potential misuse as a performance indicator', *Public Administration Review*, 39(1), 46–52.
- Strack, F., Martin, L., Schwarz, N. (1988) 'Priming and Communication - Social Determinants of Information Use in Judgments of Life Satisfaction', *European Journal of Social Psychology*, 18(5), 429–442.
- Stritch, J.M., Pedersen, M.J., Taggart, G. (2017) 'The Opportunities and Limitations of Using Mechanical Turk (MTurk) in Public Administration and Management Scholarship', *International Public Management Journal*, (Online version), 1–47.
- Tetlock, P.E. (1983) 'Accountability and the Perseverance of First Impressions', *Social Psychology Quarterly*, 46(4), 285–292.
- Tourangeau, R., Rasinski, K.A. (1988) 'Cognitive processes underlying context effects in attitude measurement', *Psychological Bulletin*, 103(3), 299–314.
- Tourangeau, R., Rasinski, K.A., Bradburn, N., D'Andrade, R. (1989) 'Belief accessibility and context effects in attitude measurement', *Journal of Experimental Social Psychology*, 25(5), 401–421.
- Van de Walle, S., Van Ryzin, G.G. (2011) 'The order of questions in a survey on citizen satisfaction with public services: Lessons from a split-ballot experiment', *Public Administration*, 89(4), 1436–1450.
- Van Ryzin, G.G. (2006) 'Testing the expectancy disconfirmation model of citizen satisfaction with local government', *Journal of Public Administration Research and Theory*, 16(4), 599.
- Weitzer, R., Tuch, S.A. (2005) 'Determinants of Public Satisfaction with the Police', *Police Quarterly*, 8(3), 279–297.
- Young, A.I., Fazio, R.H. (2013) 'Attitude accessibility as a determinant of object construal and evaluation', *Journal of Experimental Social Psychology*, 49(3), 404–418.
- Zabriskie, R.B., Ward, P.J. (2013) 'Satisfaction With Family Life Scale', *Marriage & Family Review*, 49(5), 446–463.