Working paper for discussion at the INOGOV experiments workshop in Helsinki: 11-13 March 2015

B.McFadgen, Institute for Environmental Studies, VU University, Amsterdam, the Netherlands.

Are we *Learning-by-Doing* Policy Experiments?

The use of experimentation in Dutch climate adaptation and the effect of experiment design on learning in a policy network.

DRAFT PAPER. PLEASE DO NOT CITE WITHOUT WARNING!

ABSTRACT

Studies on experimentation for climate governance are steadily growing in number as governments and communities seek to address the wicked problem of climate change. However, most studies focus on mitigation efforts, with far less attention spent on how society is addressing increasing adaptation needs. Moreover, despite experiments being the integral part of adaptive management's learning-by-doing approach little is understood of what learning is produced by experiments and how it is generated. The understanding of the relationship between the two concepts is overly simplistic and only a few empirical studies have been undertaken to explain their dynamics (e.g. Armitage et al. 2008; Farrelly & Brown 2011; van der Heijden 2014). Therefore, this paper seeks to explain how learning is produced by experimentation with a focus on external learning- that within the surrounding policy network; using proxy indicators for learning: how credible, salient, and legitimate political decision makers perceive the evidence to be.

It uses data from 17 policy-relevant experiments that produced or are still producing evidence for their surrounding policy network. The experiments are grouped in a three-way typology of ideal types drawn from the science-policy interface literature: the technocratic, boundary, and advocacy ideal types. Based on the literature, three hypotheses were constructed regarding how the types produce different levels of policy learning: technocratic experiments will produce the most credible evidence, the least salient, and moderately legitimate; advocacy experiments will produce the most salient evidence but the least credible and legitimate; and boundary experiments will produce legitimate evidence that is also perceived to be moderately salient but not very credible. In order to test the hypotheses, the surrounding policy network of each case was surveyed for their perceptions on the experiment's evidence. From a total response count of 164, 70 full responses were given for 14 experiments from three institutional levels. The results show that on the whole the evidence from experiments is perceived favourably, with positive average scores for all three types. However, the hypotheses were partially met, as surprisingly technocratic experiments were perceived to be the least credible, salient, and legitimate. This raises questions about some of the apparent

trade-offs that must be made when designing science-policy interfaces, as the data indicates that one particular institutional design can almost meet all three learning attributes.

INTRODUCTION

Policy experiments are a phenomenon that is receiving increasing attention as policy makers seek innovative ways to solve complex policy issues. Policy making is a dynamic craft; it governs society within an increasingly complex social-ecological system and ideally, policy decisions will be better decisions if they are based on scientific evidence (Pawson 2006; Sanderson 2009). Brimming with characteristics such as complexity, variability, non-reducibility, and a collective quality (Dryzek 1987), social-ecological systems require governance different to that pushing for narrowly focused optimization (Walker 2012). One suggested method to increase adaptive capacity in governance arrangements is adaptive management, which emphasizes the inherent uncertainty and complexity in social ecological systems and advocates for a learning-by-doing approach (Walters & Holling 1990; Lee 1999; Armitage et al. 2008; Huitema et al. 2009). Here, innovative management practices are cast as experiments and tested for their effects through monitoring and evaluation programmes, as a sort of *ex ante* evaluation process. Based on the results, subsequent adjustments to the approach are made. Support for an experimental approach to governance also stems from Campbell's 'Experimenting Society' (1998). The underlying premise of the experimenting society is that we are essentially ignorant of how to solve the world's problems and need to create and critique policy-relevant knowledge through policy experimentation, which would lead to new forms of public action (Dunn 1998).

Despite their strong conviction that experimentation and learning will enable better political decisions, neither adaptive management nor the experimenting society address the political repercussions of experimentation and seem to take for granted that learning will result. For example, one factor of adaptive management is that it expects the results of its experiments to be automatically used in iterations of future policy; with no consideration for the fact evidence is but one consideration for elites making policy decisions (Vedung 1998; Sanderson 2009). Likewise in the experimenting society, despite evaluation being central to experimentation, it is often found to be politically difficult or impossible to undertake (Peters 1998:133). Evaluation of experiment effects on policy is also empirically lacking. Greenberg et al (2003) conducted a thorough analysis on the political impacts of five experiments, and Milo and Lezaun (2006) assessed two experiments in regulation for their political impacts, but no studies have explicitly captured the extent of learning emanating from experiments and rarely is policy learning examined from an evaluation perspective (exceptions are Sanderson 2002; Teirlinck et al. 2013).

In conclusion, it is the purpose of this paper to pour some empirical analysis into the academic absence that is the relationship between experiments and learning in the policy network, which we attempt to do by addressing the following research questions:

1. What is the significance of a policy experiment to a policy network? How innovative and important to policy are experiments really?

2. To what extent does a policy network learn from evidence produced by a policy experiment and what kinds of learning do we observe?

3. What explains the difference in learning results among multiple experiments?

To address these questions, the paper is set out as follows: first, we outline how we understand the concept of experimentation and how we plan to measure policy learning. Next, we present an analytical framework that we use to explain the variation in learning effects found in a group of policy experiments. The framework is built around theoretical premises in the policy sciences literature and focuses on the use of science in policy making. It understands experiments as temporary science-policy interfaces that have three 'ideal type' institutional arrangements: the technocratic, boundary, and advocacy type. Based on factors found in the literature, we hypothesise that these ideal types produce different learning effects and we explain how we measure learning by assessing how credible, salient, and legitimate an experiment's evidence is perceived to be by decision makers in an experiment's surrounding policy network. This section is followed by an explanation of data collection and survey methods used. The survey data is analysed to assess the extent of this relationship. Finally, we discuss the main findings of and limitations to this research.

THEORY

Use of experimentation in policy making: Reforms on trial

Policy experimentation has an extensive political and academic history, with analyses conducted since the 1960s when the idea of the Big Society took shape. DT Campbell was one of the first to kick against what he saw as policy decisions being taken without the risk of criticism or failure and he advocated the use of policy evaluation using experimental and quasi-experimental approaches (Dunn 1998; Sanderson 2009). The method gained traction and during the following decades experimental interventions were conducted in an attempt to improve economic, health, and education policy, particularly in the US and UK. Criticisms of the use of experiments included ethical issues of treating citizens as subjects, their use to delay action and political commitment, and the belief that the complexities of the social world could never be understood and managed through such a limiting prism (Fischer 1995; Sanderson 2002; Greenberg et al. 2003). The concept subsequently lost support, but enjoyed a revival with the emergence of the adaptive management approach in environmental governance (Walters & Holling 1990; Lee 1999). Here, experimentation is expected to provide reliable evidence of whether new management interventions worked, as well as provide a vehicle for incorporating a broad range of actors and new ideas into the policy process through the implementation of shadow networks (Meijerink & Huitema 2007). Now, the concept

also has a place in the policy innovations literature, where they maintain an evaluative function as well as being epicentres of new, innovative forms of governance (Jordan & Huitema 2014).

These different uses of experimentation highlight varying understandings of the concept, but have in common the positive characteristic of being flexible- in that experiments are never implemented on a full scale and are reversible (Tassey 2014). A useful definition that captures the important characteristics of policy experimentation is: *a temporary, controlled field-trial of a policy-relevant innovation that produces evidence for subsequent policy decisions.* This definition is considerably narrower than that of a related concept, the pilot project, which also has an early evaluative function but with multiple purposes: to begin roll-out of a new policy (pioneer), demonstrate successful implementation of a policy (demonstrator), operationalise a policy to overcome barriers (trailblazer), as well as experiment with new approaches that have uncertain effects (policy trial/experiment) (Ettelt et al. 2014). The experiment definition also differs from the governance innovations literature in that it emphasises the testing component (compared to Hoffman 2011 where experimentation is recognised for its novelty function outside established policy order).

In essence, we consider experimentation as the appraisal of a policy relevant innovation in practice: without appraisal you are demonstrating a new initiative, without innovation you are evaluating established ideas. These descriptors go some way to explain why as a method of developing new policy approaches, experimentation is actually quite uncommon (Peters 1998; Gunderson 1999; Sanderson 2009). Even if governments do occasionally attempt institutional reform, the evaluation of outcomes is lacking (Campbell 1998). Moreover, innovation itself may be seen as limited, with policy change mostly occurring incrementally within existing programmes and not in the bursts that you would expect if radical innovation was common (Vedung 1998:193).

Policy learning

Learning is a goal of the experimenting society (Campbell 1998). It is implicit in the presence of a monitoring and evaluation framework, which declares intent to experiment and learn, compared to the implementation of an innovative and creative policy solution with the intention of getting it right the first time (Peters 1998). So, learning may be produced, but how can we measure the extent of it? The learning literature presents two options. First, in line with Hall (1993), who grades policy learning from minor policy amendments to major policy shifts, we can seize the moment of learning as when policies are seen to have changed (see also Bennett and Howlett 1992). This perspective is utilised by Owens (2010) to explain how expert knowledge influenced policy formation in UK climate policy, and learning can thus be explained by the experiment producing change within the policy system. However, several caveats must be met to use this approach. First, Weiss (1977) notes how the effects of experiments on policy networks are likely to be indirect and protracted, so measuring them directly is difficult. Policy changes are really only detectable over a matter of years, Sabatier states a minimum of 10 years is appropriate, so only older experiments could be assessed. Also, the complexities of policy formation and reformation mean isolating

variables that explain change is a huge task, and a mammoth task for a comparative study of multiple experiments. Another way to assess subsequent policy learning from experiments is to measure the change in an individual playing a role in the network. More in line with Sabatier's definition of learning as an enduring alteration of thought or behaviour (1988) we can focus on an individual's recorded change in understanding, based on their experience of the experiment. It may not be as thorough an analysis as tracking the influence of the experiment's evidence on long term policy changes, but it allows for young experiments to be assessed and in higher numbers.

This is the path we take in this research; gauging the impact of an experiment on its policy network by assessing the change in perception of policy decision makers who may have been influenced by the experiment's evidence. One approach to measuring how effective experimental evidence is for policy is to engage Cash *et al*.'s 2002 science-policy boundary model. The credibility, salience, and legitimacy of a science-policy interface can be seen as important determinants of the effectiveness of its evidence (Sarkki et al. 2014). *Credibility* refers to the degree to which policy makers consider the findings of the experiment authoritative and believable, and to the degree in which they trust the outcomes. *Salience* refers to the relevance of the experiment findings at a certain moment in time. *Legitimacy* refers to the degree to which an information producing process was fair and whether it considered appropriate values, concerns, and perspectives of different actors (Cash *et al*. 2002). These indicators are chosen here because they are well established in the literature, and because they make good proxy learning indicators. It is expected that fulfilling these criteria better will lead to higher learning effects in the longer run.

Analytical Framework

An experiment is an avenue for connecting science (knowledge) to policy (action). In the literature, science-policy interfaces ('SPI') are defined as: "social processes which encompass relations between scientists and other actors in the policy process, and which allows for exchanges, co-evolution, and joint construction of knowledge with the aim of enriching decision-making and/or research" (Sarkki et al. 2014). Experiments can be seen then as a temporary site where science and policy can 'engage in elaborate and productive interplay' (Munaretto and Huitema 2012). Drawing on policy sciences (e.g. Dryzek 1987; Owens et al. 2004; Sanderson 2009) and SPI (Pielke 2007) literature, we construct a model that reflects how experiment evidence is constructed and used, and draw assumptions about how experiments might be designed under this model. The ideal types encapsulate three perspectives of the role of science in policy making: the technocratic, boundary, and advocacy experiments; and the various configurations of institutional rules that underlie the ideal types form the foundation of our learning analysis in the following section. The rules stem from Ostrom's Institutional Analysis and Development framework (2005) and are those of an action situation that determine who is involved and who is excluded (boundary rules), how tasks and responsibilities are distributed (choice rules), what types of information are distributed, how regularly, and to whom (information rules), the extent of buy-in by participants (pay-off rules); and how decisions are made.

It is contended in this paper that the types vary in the extent to which they impact the perception of political decision makers relevant to an experiment. Based on the SPI literature we argue that an experiment's rule configuration, and thus its ideal type, can be used to explain how credible, salient, and legitimate the policy network perceives the experiment's evidence. The idea that a SPI's design choices affect its evidence has support. Sarkki et al (2014) attempt to understand trade-offs made by SPIs and identify several management and design choices which they claim improve a SPI's effectiveness, and several of the factors they claim influence the success of a SPI are summarised in table 1. How these design choices shape each ideal type is explained in the following section.

Learning indicators	**Corresponding rule**	**Design choice to improve learning**
Credibility	Boundary rule	Isolating experts, neutrality or bias in position
	Boundary rule	Limiting participation of non-state actors
	Info rule	Only utilising scientific knowledge
	Info rule	Distributing knowledge to everyone
	Info rule	Open and transparent distribution
Salience	Info rule	Utilisation of place-based, non-scientific knowledge
	Info rule	Effective and timely distribution of information
	Boundary rule	Involving actors of like mind
	Boundary rule	Engaging decision makers
Legitimacy	Boundary rule	Increased inclusiveness
	Position rule	Openness to new participants
	Position rule	Facilitation, conflict management
	Authority rule	Designating power to all participants
	Info rule	Open and transparent distribution
	Aggregation rule	Consensus based decision making

Table 1: A summary of how the proxy learning indicators relate to the institutional rules.

Technocratic ideal type

The technocratic experiment resembles the technical-rational model of policy decision making, where an expert elite generates scientific knowledge for policy decisions (Owens et al. 2004; Fischer 2007). In this arrangement, the experiment produces scientific information with little or no connection to the policy process until the end, when the results are presented to decision makers. Scientists thus play a vital, but objective and disconnected, role in politics as 'science arbiters' (Pielke Jr. 2007). These expert actors are the initiators and sole participants of a technocratic experiment and maintain control over its design,

monitoring, and evaluation. Due to political disagreement, policy actors commission the experiment in order to produce factual evidence, so they fund the project and they develop an action theory and set the policy goals that the experiment needs to fit in in advance. However, they are disconnected from the project itself and have no decision authority. Scientific knowledge is the only type of information valued and generated by the experiment and fact finding occurs within the parameters of the goals previously set. This arrangement separates science from policy decision making and helps reinforce the view that science is impartial to politics, which upholds the scientific integrity of the evidence but may compromise its policy relevance.

Based on our understanding of Cash's typology, we argue that these design features influence the perception of the produced evidence in terms of its credibility, salience, and legitimacy. First, we would expect the experiment evidence to be considered highly credible due to the emphasis on independent scientific methods and expertise, as well as the open and transparent transmission of scientific information to participants. However, limiting participation to expert actors only and excluding discussion on different perspectives means the experiment is less likely to produce knowledge that resonates with the needs of policy makers, reducing the possibility of evidence being considered salient. The closed character of the experiment makes the legitimacy of the results questionable as the research question, data gathering process, and report writing has not involved stakeholder groups or ordinary citizens and might not address arguments they consider important.

Boundary ideal type

A boundary experiment is one in which policy actors open up the policy process to any actor, state or non-state who has a desire to influence policy making. The role of the scientist resembles the 'honest broker of policy alternatives' (Pielke Jr. 2007), where they engage with the policy process and develop policy solutions in accordance with multiple value-perspectives. A boundary experiment is initiated by a collaboration of actors and the production of scientific knowledge is supplemented by multiple knowledge systems- relevant contextual, lay and traditional forms of knowledge, which are considered of equal value (Koetz et al 2012). This policy relevant knowledge is also subject to an extended societal peer review (Funtowicz and Ravetz, 1993) as non-state actors have influence and decision power over the experiment's design, monitoring, and evaluation. Discussion over goal setting is high in a boundary experiment and there is reflection on whether the experiment adheres to acceptable societal aims. Deliberative practices are encouraged with transparent information transmission, open dialogue, and regular communication among participants. Ideally, this engagement will allow different interpretations of the policy problem to emerge that build into a common consensus on the most appropriate course of action (Dryzek 1987).

These design features are expected to produce quite different learning outcomes to the technocratic type. Such wide boundary settings ensure that non-state actors have access to policy making where they can influence how a public policy problem is solved. We would expect this to mean the experiment evidence would be perceived as very legitimate, as the inclusion of different perspectives increases the chance that

the evidence resonates with societal needs (Hegger et al 2012). Moreover, open and transparent information transmission between participants allows for the 'extended peer review' of experiment evidence by a range of actors, rendering the information produced more relevant to policy and more legitimate. The inclusion of different knowledge types will dilute the sense of independent and reliable knowledge being produced; thereby lowering the perception of credibility, but this design would open up debates about what knowledge is relevant to addressing pressing policy needs.

Advocacy ideal type

An advocacy experiment produces evidence that steers policy towards a pre-defined position (Pielke Jr. 2007). It is organized by policy makers and populated by dominant, traditional actors (Hoogma et al. 2002), which may make it a diverse group but they are all part of the same advocacy coalition. Although appearing neutral, the experiment supports particular outcomes as participants must be invited and outsiders are barred from gaining access (Owens et al. 2004). Involvement is restricted by certain conditions and those with contrasting expectations are excluded. A steering group of dominant participants control the design, monitoring and evaluation procedures, reinforcing existing structures of power. Within the group, only the dominant participants discuss and shape goals through the use of a facilitator so prevailing norms are protected, which also limits the generation of new ideas. Sensitive information is not shared so information distribution and openness tends to be low within the wider group as well as with outsiders. Advocacy experiments are producers of *policy-based* evidence, where despite it producing new knowledge about a policy idea, the experiment is designed so that it suppresses unintended or unflattering results and only promotes the promised benefits.

The expected learning patterns are as follows: credibility is undermined by including policy and non-state actors in the experiment along with expert actors, as does the production of practical knowledge alongside scientific knowledge. Moreover, if it is noticed, the process of cherry-picking information affects the reliability of the results. In the attempt to show there is support for a particular proposal, the initiator blocks participation by critical actors and thereby undermines their concerns, reducing fairness and the perceived legitimacy of the project. However, the salience of the findings may be perceived of as high when the experiment acts as a means for keeping an idea alive (Greenberg et al. 2003), and outcomes are presented when the time is right- carefully gauged and engineered by the policy actors involved.

In summary, the three types are expected to produce different levels of learning effects. Appendix A clearly illustrates the distinctions between the types and their rule settings. Table 2 summarises the expectations sketched above into three tentative hypotheses:

	Credibility	**Salience**	**Legitimacy**
H1:technocratic type	Highest	Lowest	Middle
H2: boundary type	Middle	Middle	Highest
H3: advocacy type	Lowest	Highest	Lowest

Table 2: Summary of hypotheses that link ideal types to specific learning indicators.

It has been suggested that it is difficult to maximize all three criteria (Cash et al. 2002), and attempts to improve one criterion might actually lead to lower scores for the other criteria. For example, by engaging policy makers you may improve salience because of the increased probability that the right questions are being asked. However, credibility is lowered if science is seen to be biased by the policy process. Likewise, limiting access from outside may increase credibility but decrease legitimacy. These tensions are attractive for the current research as the various types of experiments may be designed to score well on different criteria and thus lead to learning effects amongst decision makers in diverging ways.

METHODS

Case selection

In order to utilise the framework, we need a set of policy experiments. Based on the policy science and adaptive management literatures, we identified six criteria for isolating experiment cases from a broader set of 147 innovative pilot cases related to adaptation: whether the project was testing for effects; whether it was innovative with uncertain outcomes; whether it had policy relevance; whether there was state involvement; whether it was eliciting an ecosystem response, and whether it was relevant to climate adaptation. These six are summarised in table 3) and elaborated on in turn in Appendix B. 18 cases were selected as meeting all six criteria, the most uncommon criteria being monitoring and evaluation framework and climate adaptation relevance. The cases have different spatial and temporal scales and deal with different problems; however, they are comparable due to their meeting the stringent conditions.

Criteria	**Indicators**
Testing for effects	Monitoring and evaluation framework in place
Innovative	Long-term alteration in policy or management practice
Policy relevance	Testing manifestation of new policy concept or approach
State involvement	Either as initiator or a minor party
Ecosystem response	Intervention straddled the social-ecological system
Climate adaptation	Any bearing on adaptation planning at all

Table 3: Sets out the six criteria and associated indicators used to identify policy experiments in Dutch climate adaptation.

Experiment cases

Climate adaptation is increasingly understood as a matter of urgency in a lowland country such as the Netherlands, as it is particularly vulnerable to sea-level rise, flooding, salt-water intrusion, fresh water availability, and increased drought. The 18 experiment cases in Dutch climate adaptation tested policy innovations in coastal defence, water availability, multi-functional land use, water variability, and dike management. The names of the experiments are not given to honour confidentiality. They date between

1997- 2012 and almost half are ongoing. Ongoing cases were included in the analysis if they have passed at least one substantial evaluation phase.

Based on previous survey data and a document analysis, each case was assessed and assigned an ideal type. In order to do this, they were judged on a set of indicators based on the institutional rules. Each indicator has a setting for the three ideal types. The indicators and specific rule settings are contained in Appendix A. In order to determine what type an experiment fell into, the cases were assessed against each indicator and were labelled the ideal type that was most common. For example, out of 14 indicators, experiment 2 had one score for a technocratic setting, 9 scores for boundary, and 4 scores for an advocacy setting; thus it was classified as a boundary experiment. The assessment concludes that for the 18 cases, five experiments met the technocratic definition, six were boundary experiments, and seven were advocacy experiments. It was uncommon for cases to fall absolutely into one type, for the ideal types are just theoretical constructions on which to base reality. However, out of 14 indicators one ideal type typically emerged as the dominant type, and each case was duly assigned. For this research the case number fell to 17 because one experiment was omitted due to it being a national experiment and therefore relevant to all institutions throughout the country.

Data collection- learning

In order to measure the influence of each experiment on its surrounding policy network, we conducted a desktop search to identify decision makers at each relevant institutional level: being the municipality, the water board, the province, and the ministerial level (where appropriate). Using random sampling we produced a list of 40 respondents for each experiment, and each respondent was emailed the survey link. If a case did not generate responses, other respondents were identified on the list. To encourage participation and prove the legitimacy of the survey, we included in the email an endorsement from the President of the Dutch Union of Water boards, Mr. Peter Glas. The initial email was followed by two reminders.

Survey design

The 40 respondents were first asked whether they had heard of the experiment conducted in their area. If they responded yes, the respondents were asked a series of closed questions and statements to gauge their perception on how credible, salient, and legitimate they thought the experiment was. If they did not know of the experiment they were directed to the end of the survey. Other questions included: how relevant they thought the responses were to their policy decisions, how innovative they thought the project was, what the rate of policy change experienced by their organisation was, and for what reason they thought the experiment was organised.

Data Analysis

The survey results were analysed using statistical data-analysis software (SPSS 21). Basic descriptive statistics (frequency tables, cross-tabulations), Kruskal-Wallis tests to gauge differences in distribution and ANOVA (analysis of variance) tests were used to assess the statistical significance of variation.

RESULTS

This section describes the results of the survey, the characteristics and significance of the cases, and the extent of policy learning from each case.

From the sampling a total of 164 responses were received, of which 60% had heard of the experiment relevant to their area. Completed surveys were rarer (N=79). Appendix C sets out, for each case, the number of initial responses received, how many of those respondents had knowledge of the experiment in question, and how many went on to complete a survey in full. Appendix C also notes for each case which institutions responded: municipality (responded in 10 cases); water authority (16); province (10); ministry (2). The survey was also sent to decision makers at the enforcement arm of the Ministry of Infrastructure and Environment, Rijkswaterstaat (Department for Waterways and Public Works) if this institution was involved in a case. As shown in table 4, the water authorities were most heavily represented, possibly because of the endorsement by their head, as well as the issues being most relevant to their organisations (see Appendix B).

Institution	***# responses***	***Percent knew case***	***Percent completed survey***
Municipality	37 (23%)	57	36
Water Authority	68 (42%)	73	66
Province	54 (34%)	43	38
Ministry	2 (1%)	100	50
TOTAL	161[1]	62%	47.5%

Table 4: Extent of responses from individual institutions

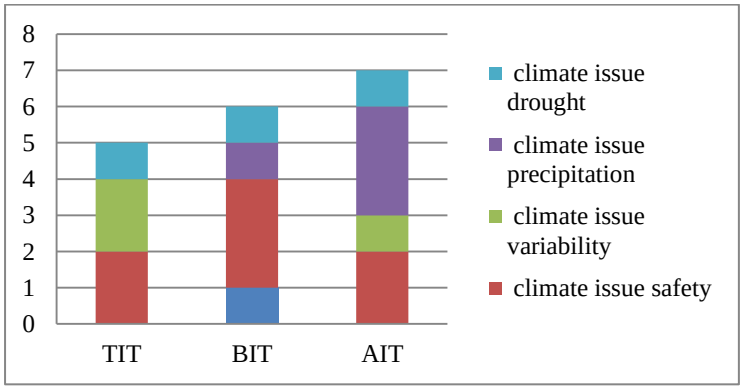
Experiments

Characteristics of the experiments

Experiments differed according to what climate issue they tackled and the main reasons they were implemented. The experiments dealt with a range of climate issues, from safety against sea level rise, increased precipitation, water variability, drought, and salinity, and these differed somewhat between ideal

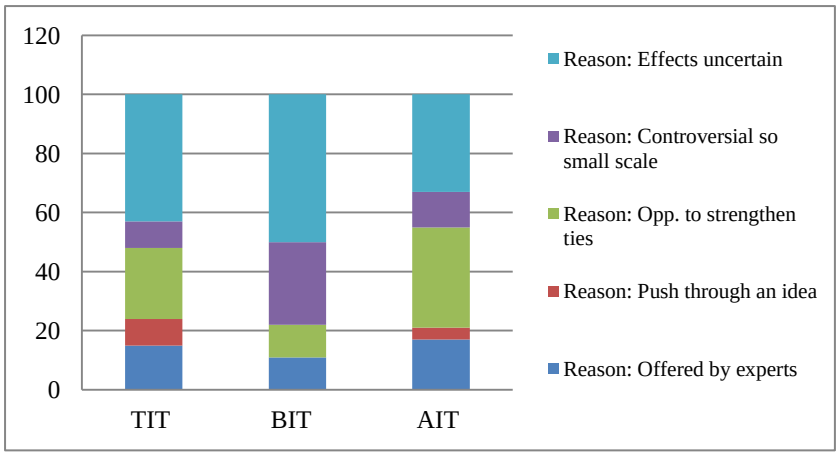
[1] Three respondents did not give their institutional affiliation.

types (graph 1). Safety and precipitation- usually testing multi-functional land use approaches- were most common. Although the cases varied in issue, what they did have in common was their affiliation with water. This is explained by the overlaying of climate adaptation measures predominantly atop of the established water management issues, thereby allowing the Dutch government to claim they are responding to climate adaptation without changing their incumbent institutional structure. This is understandable since the Dutch are particularly vulnerable to water related climate events such as sea level rise, salt water intrusion, increased precipitation and drought (Biesbroek et al. 2011) and other adaptation related responses- such as land-use planning and agriculture- are coupled to water concerns (e.g. multi-functional land use, self-sufficient farming).



Graph 1: Number of experiments for each climate issue, divided by ideal types

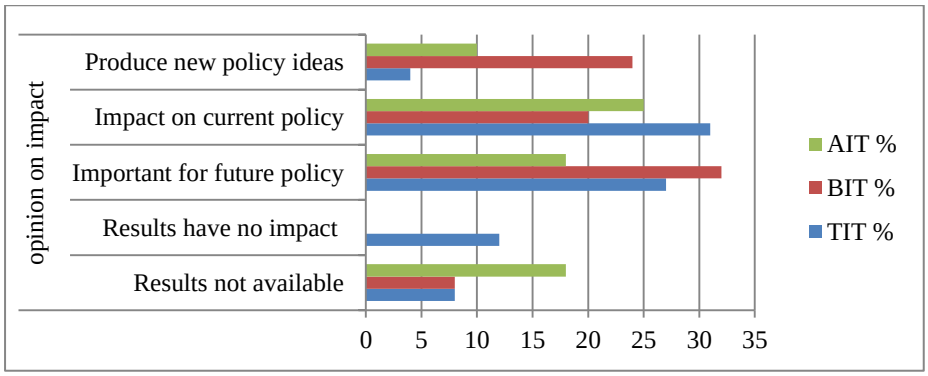
The main reason for organising an experiment was seen to be measuring a new approach for unknown effects, followed by the opportunity to strengthen ties among actors (graph 2). The main reason for each type is to test for unknown effects; although the graph shows that advocacy experiments are most seen to strengthen ties, and boundary types are seen to be controversial and needing to be tested on a small scale.



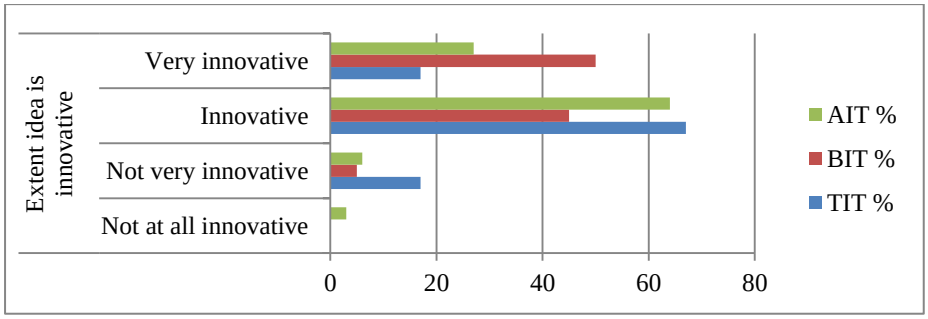
Graph 2: Perceived reasons for conducting experiments grouped according to ideal type.

Significance of experiments

To assess the significance of the cases, we asked decision makers what they thought the impact of experiment findings would be on their policy choices. Graph 3 shows that boundary experiments are considered to have the most impact on policy ideas, and in general experiments are expected to have impact on current and future policy. Of those who had an opinion, very few decision makers believed that experiment evidence would have no impact on policy decisions. We also asked decision makers how innovative they thought the experiments were, with only 10% stating they thought their case was not very or not at all innovative (graph 4). Boundary types are even significantly perceived as more innovative than the other two (Kruskal-Wallis test for distribution across ideal types: $p<0.05 = 0.046$).



Graph 3: How the decision makers' opinion on experiment impact differs according to ideal type.



Graph 4: How the ideal types differ according to how innovative they are seen to be.

Policy learning

The following section presents the survey data results for each aspect of policy learning we measured. A minimum of four full responses was the cut-off and due to the lack of responses from cases 7, 8 and 9 they were removed from the analysis. We will revisit these cases in the discussion, as it is of interest that despite a high initial response rate, cases 7 and 9 were not heard of by relevant institutions, but for the

learning assessment we use a total of 14 cases- five technocratic, four boundary, and five advocacy experiments.

Credibility

In order to assess how decision makers perceive the scientific adequacy of the technical evidence and arguments produced by each experiment, we asked them a series of seven questions. The questions gauge a decision maker's perception of data quality, their trust in the experts, and the standard of the conclusions, which are all composite factors that can be put together and used to measure credibility as a variable. The questions were answered on a scale ranging from: (1) no certainly not; (2) not really; (3) neutral; (4) somewhat; (5) certainly. Reliability of each composite measure was determined by computing Cronbach's alpha (α). The factors had a high reliability score (α= 0.92), well over the 0.7 needed to justify the aggregation into one variable (De Vaus 2002). The entire sample averaged 4.1 with a range from 2-5, which means on average decision makers considered the evidence produced by the experiments as somewhat credible (table 5). The response means differ slightly among different institutional levels, with municipal decision makers scoring the experiments slightly lower than the others, but there is no significant difference in the scores between parties.

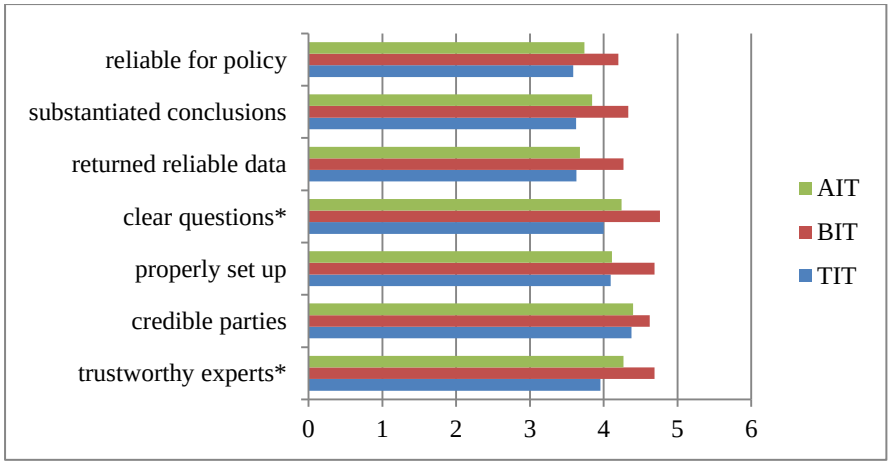
	N	Minimum	Maximum	Mean	S.D.	Significance between cases
Credibility score	70	2	5	4.1	0.64	Yes ($p<0.05= 0.02$)

Table 5: Summary of respondent's scores for the variable **credibility**.

When the average credibility scores are divided up among the ideal types we see that contrary to our hypothesis, the technocratic experiment actually scores lowest out of the three types. Further, an ANOVA test reveals that the average of the scores across the three types (table 6) differ significantly across types ($p<0.05 = 0.033$) with the scores from the boundary experiments being significantly higher than those measured in the technocratic experiments ($p<0.05 = 0.026$). Graph 5 shows the distribution of scores for each of the types over the seven measured indicators. Running a kruskal-wallis test shows that for the indicators "trustworthy experts" and "clear questions" the results are significantly different for each ideal type, so these indicators are where the technocratic types are the weakest.

Ideal Type	**N**	**Mean**
Technocratic	24	3.9
Boundary	17	4.5
Advocacy	29	4.1

Table 6: Comparing ideal types mean scores for credibility



Graph 5: how the ideal types differ for each question measuring credibility.

Salience

In order to assess how decision makers perceive the relevance of the evidence to the needs of policy decision makers, we asked them to respond to a series of nine statements. The statements gauge a decision maker's perception of whether the findings filled a knowledge gap for policy makers, whether the experiment was a matter of public interest, whether the results were adequately communicated to policy actors, and whether the evidence created an opportunity to renew policy. These composite factors can also be put together and used to measure salience as a variable. The questions were answered on a scale ranging from: (1) strongly disagree; (2) disagree; (3) neutral; (4) agree; (5) strongly agree. Reliability of each composite measure was determined by computing Cronbach's alpha (α). The factors had a good reliability score (α= 0.84), again over the 0.7 needed to justify the aggregation into one variable.

The results show that on average, survey participants responded favourably to the statements (mean= 3.6) meaning the decision makers on the whole found experimental evidence relevant to policy making (table 7). The different institutions grade the experiments lower for salience than for credibility, although this time municipalities are slightly more favourable than the others.

	N	Minimum	Maximum	Mean	S.D.	Significance between cases
Salience score	72	2.22	5	3.6	0.6	Yes ($p<0.05= 0.005$)

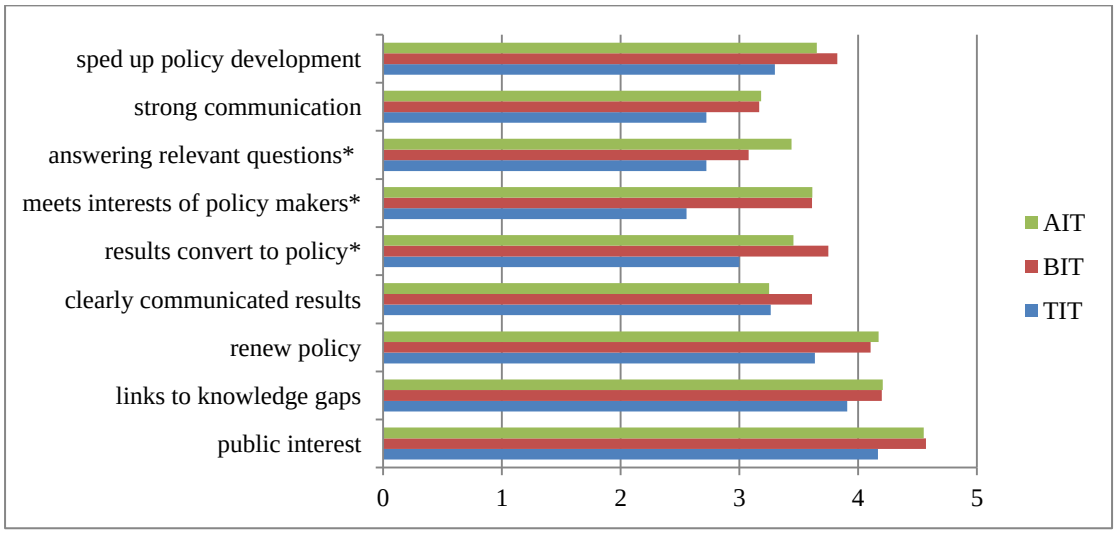
Table 7: Summary of scores for the variable salience

When comparing the scores for each ideal type (table 8), we see our hypotheses are partially met, in that the technocratic type is significantly less salient (p=0.002). However, contrary to expectations the boundary type experiments scored exactly the same as the advocacy experiments. By reviewing graph 6 we can see how each type scores on the individual questions, with the boundary type scoring higher on statements about "how well the results were communicated", whether "the experiment was within the public interest" (p=0.045), and "how well the results converted directly to policy" (p=0.022). In contrast,

advocacy experiments scored best on statements regarding "answers relevant questions" (p=0.017) and "provides chances to renew policy". The statements with a p-value are those that were found to be significant after conducting a kruskal-wallis test, illustrating again where the technocratic experiments are weakest. Technocratic experiments are perceived of as not salient in three indicators that dip below the middle score of three, confirming the assumption that separating science and policy reduces communication, question relevance, and renders the results of little use of policy makers.

Ideal Type	**N**	**Mean**
Technocratic	24	3.3
Boundary	19	3.8
Advocacy	29	3.8
Total	72	

Table 8: Comparing ideal types mean scores for salience



Graph 6: how the ideal types differ for each statement measuring salience.

Legitimacy

The last variable to assess is legitimacy. In order to assess how decision makers perceive how fair and balanced the experiment process was, we asked them to respond to five questions. The questions gauge a decision maker's opinion on whether the experiment included all relevant parties from the area, whether perspectives were treated respectfully, how transparent the process was, and whether the goals of the experiment were in line with community values. These composite factors were put together and used to measure legitimacy as a variable. The questions were answered on a scale ranging from: (1) no certainly not; (2) not really; (3) neutral; (4) somewhat; (5) certainly. The factors had a high Cronbach's alpha score

(α= 0.88), over the 0.7 needed to justify the aggregation into one variable (De Vaus 2002). The results show that on average, survey participants responded favourably to the statements (mean= 3.8) meaning the decision makers found the process of experimentation legitimate (table 9). The institution scores vary slightly, with the provinces this time scoring experiments the highest for their legitimacy, but the scores do not vary significantly.

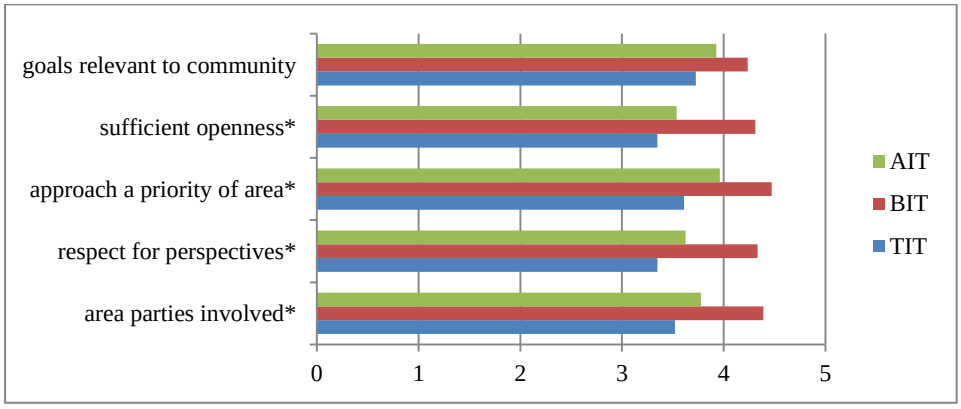
	N	Minimum	Maximum	Mean	S.D.	Significance between cases
Legitimacy score	69	1	5	3.8	0.76	Yes ($p<0.05= 0.009$)

Table 9: Summary of scores for the variable legitimacy.

When it comes to evaluating the differences in ideal types, again the boundary experiments come out on top. Table 10 shows that there is a marked difference in the averages, with boundary experiments significantly more legitimate than technocratic (p=0.002) and advocacy (p=0.046) experiments. When we assess the questions more closely (graph 7), we see that despite their lower scores technocratic and advocacy experiments still score above neutral in all areas, but are still far behind the boundary experiment in nearly all indicators.

Ideal Type	**N**	**Mean**
Technocratic	24	3.5
Boundary	17	4.3
Advocacy	28	3.8
Total	69	

Table 10: Comparing ideal types mean scores for legitimacy



Graph 7: how the ideal types differ for each statement measuring legitimacy.

DISCUSSION AND CONCLUSION

What can the results tell us about how the design of an experiment affects a decision maker's perception of the evidence? First, design seems to matter, and the significance of scores between ideal types indicates that we are on to something by explaining learning with this model. However, the assumptions we drew from the literature about how design impacts learning were not mirrored by the empirical findings. Technocratic experiments scored well on most indicators, but still consistently lower than the other two types. Salience was expectedly low, but a technocratic experiment being least credible was a surprise. To decision makers at least, isolating experts from policy and limiting participation makes an experiment less credible, not more. It mattered less that these experiments had more open and transparent information channels than the advocacy experiments. The results also highlight a contradiction in the literature. At its inception, credible evidence was defined as the perceived scientific adequacy of a SPI's technical evidence and arguments (Cash et al. 2003). Generalizable, scientific knowledge was considered most plausible and accurate. However, in the theoretical development of these indicators, credibility has essentially been broadened to include place-based knowledge in science (e.g. Hegger et al. 2012). For experiments at least, when a broad set of actors are seen to contribute contextual, practical knowledge, this place-based knowledge counts for more than scientifically defensible knowledge even with a lack of transparency and information distribution between the participants. Although it must be noted that when transparency and distribution are high, the perception of credibility is sky-high- as shown by the success of boundary experiments.

Our results for salience were also partly surprising, as boundary experiments managed the same moderate score as advocacy experiments. There is some doubt whether boundary experiments can always produce salient knowledge, especially if decision makers are responding to cues other than societal norms (e.g. international, economic, political influences) but in these instances they were seen produce relevant results. The salience of the sample was lower than the other indicators of learning, perhaps indicating the difficulty for experiments to maintain their connection to policy. Finally, from reading the scores for legitimacy it is plainly clear how increased inclusiveness and openness improves this attribute. Boundary experiments were perceived of as significantly more legitimate than the other two types on nearly all the legitimacy indicators. If policy makers want to be seen as meeting societal goals and creating fair policy processes, then they need to pay attention to their experimental design.

When considering the impact of the results on theory we turn to the SPI literature. Cash et al (2002) state that although the salience, credibility and legitimacy attributes are not isolated from one another in a SPI, trying to increase the likelihood of one actually causes tensions, in that bolstering one attribute will often decrease the extent of another. We will apply our findings to two main tensions (ibid.). Tension one is the design choice of whether to engage policy actors/decision makers in the experiment, thereby isolating science from the policy process. We would expect that including policy actors means the questions would be policy relevant and increase salience, but it would decrease credibility due to science being seen to be biased by political concerns. As discussed above, experiments with a broad set of actors had more

credibility than those without policy actors/minimal policy participation. Experiments furthest away from policy generated the least trust in their experts and questions were considered the least clear, maybe explained by a fear of 'science biasing policy'. The second tension is that including a broad range of actors increase relevance because the problem perception is better understood, but legitimacy is affected because some actors are involved that others think should not be. We did not see evidence of this in our cases. In fact, legitimacy was highest when there was the most actor diversity and shared authority. Our results challenge the assumption that trade-offs need to be made when designing a SPI.

The results of the survey reveal that on the whole policy experiments make a positive impression on their surrounding policy network. On average they are seen to be of high quality, and produce results that are very credible, moderately salient, and moderately legitimate, with no significant difference between scores from different institutions. The results bode well for actors looking to use the method to assess future innovations. However, that decision makers find the process of experimentation a largely positive endeavour is a useful finding, and begs the questions: why is the process not used more to develop and evaluate policy and management strategies in climate adaptation? Moreover, why had so many respondents (almost half) never heard of the projects? The first question is one for future research, but we note from our case studies that the cost and the uncertainty of risk make it hard for policy makers to swallow the idea that failure does not matter. Supporting this assumption is the number of cases that stressed the importance of building political and public support for the project before it even hits the drawing board, which is a factor we did not capture in our framework, and being explicit in the fact that if the experiment fails, the costs will be borne by the state. On reflection, it makes sense that policy actors are cautious about innovating and taking risks, since they are spending public money (Duijn 2009), although the costs of not adapting may far outweigh the costs of trying new ways to keep society's proverbial head above water.

In regards to the second question, respondents were randomly chosen but the ones with environmental portfolios were targeted first, so we can assume that experiments are being conducted in jurisdictions where relevant political decision makers are unaware of their presence. Two experiments had to be withdrawn from the learning analysis due to a lack of data despite a high response rate and one removed because there was almost no response at all. This- and the response rate in general, unfortunately reduced our case size and limited our findings. Why were decision makers not aware of three of the projects? They were different types, and two were completed, so these variables do not explain it. An election might replace decision makers, but other, older experiments were adequately responded to in the survey. Visibility and success might be relevant factors, and we failed to control for these variables. Another explanation might be that the respondents are just not interested in these sorts of projects. Pawson (2006) suggests that as one moves further up the bureaucratic chain, their appetite for evidence dwindles, so we could interpret the survey results as indicating which experiments are most visible to the political elite. How initiators of these experiments managed to catch the policy network's attention is also a question for future research.

That the design of an experiment has a strong effect on how it is perceived is a proposition derived from the literature. Other reasons could be given for the variation in learning we measured. For example, intervening variables include the extent the experiment was seen as innovative or its impact on policy. As we saw in the results, boundary experiments also scored highest on these factors. This suggests that policy actors are utilising this design for particular types of projects. The extent of constraints on design choices is another focus for future research.

In conclusion, when a policy experiment is understood as an appraisal of a policy innovation, it makes a lot of sense to analyse it as a temporary science-policy interface that produces evidence for policy action. Applying an institutional lens allows us to thoroughly test our assumptions that design matters when looking to produce learning. These assumptions are strong due to their theoretical origins, drawn from policy sciences and science and technology studies. Our research sheds light on the perception of adaptation experiments in the Dutch policy network, with interesting findings that can be used to assist policy makers in designing future experiments. However, just because the perceptions of experiment cases were generally positive in the policy network does not mean the experiments themselves were or will be adopted. Actual policy influence takes time to identify and measure, and the literature is not too positive on whether experiments really meet these expectations. Greenberg et al. (2003) discovered no instances where effects of their evaluated cases were decisive in the decision to adopt the tested policy, rather political reasons ended up trumping experiment evidence. Vreugdenhil et al. (2010) conclude also that pilot projects often have limited influence over their policy domains. So, boundary experiments guarantee the adoption of a particular policy strategy is not a finding we can claim. However, the results take us one step on the path towards understanding the kind of role experiments play in policy making.

(remainder: 2,100 words)

REFERENCES (apologies some are missing)

Armitage, D., M. Marschke, and R. Plummer. 2008. Adaptive co-management and the paradox of learning. *Global Environmental Change* **18**:86-98.

Bennett, C.J. and M. Howlett. 1992. The lessons of learning -reconciling theories of policy learning and policy change, *Policy Sciences* **25** (3) 275–294.

Campbell, D.T. 1998. *The Experimenting Society*, p35. In **Dunn, W. (ed.)** 1998. *The Experimenting Society: Essays in Honor of Donald T. Campbell. Policy Studies Review Annual volume 11*. Transaction Publishers, New Brunswick, New Jersey.

Cash, D. W. Clark, F. Alcock, N. Dickson, N. Eckley and J. Jäger. 2002. Salience, credibility, legitimacy and boundaries: Linking research, assessment and decision-making. *John F. Kennedy School of Government Harvard University Cambridge, MA, Faculty Working Paper RWP02-046.*

Cash, D. W., W. C. Clark, F. Alcock, N. M. Dickson, N. Eckley, D. H. Guston, J. Jäger, and R. B. Mitchell. 2003. Science and technology for sustainable development special feature: Knowledge systems for sustainable development. *Proc. Natl. Acad. Sci. USA* 100: 8086-8091.

Dryzek, J. 1987. *Rational Ecology. Environment and political economy.* Basil Blackwell, New York, New York, USA.

Dryzek, J. 1993. *Policy analysis and planning: from science to argument*, pp. 213-233. In **Fischer, F. and J. Forester (eds.)**. *The argumentative turn in policy analysis and planning.* Duke University Press, Durham.

Farrelly, M. and R. Brown. 2012. Rethinking urban water management: Experimentation as a way forward? *Global Environmental Change* **21**(2): 721-732.

Fischer, F. 1995. *Evaluating Public Policy.* Nelson Hall, Chicago, Illinois, USA.

Fischer, F. 2007. *Deliberative policy analysis as practical reason: Integrating Empircal and Normative Arguments*, pp.223-237. In **Fischer, F., G.J. Miller, M.S. Sidney (ed.)**. *Handbook of Public Policy Analysis: theory, politics, and methods*. Taylor & Francis Group, Florida, USA.

Funtowicz, S.O. and J.R. Ravetz. 1990. *Uncertainty and quality in science for policy*. Kluwer Academic, Dordrecht, The Netherlands.

Greenberg, D., D. Linksz, and M. Mandell. 2003. *Social experimentation and public policy making.* The Urban Institute Press, Washington, D.C., USA.

Gunderson, L. 1999. Resilience, flexibility and adaptive management - - antidotes for spurious certitude? Conservation Ecology 3(1): 7.

Hall, P.A. 1993. Policy paradigms, social learning, and the state: the case of economic policymaking in Britain. *Comparative Politics* **25(**3): 275-296.

Hegger, D., M. Lamers, A. Van Zeijl-Rozema, and C. Dieperink. 2012. Conceptualising joint knowledge production in regional climate change adaptation projects: success conditions and levers for action. *Environmental Science & Policy*, **18**:52-65.

Hoffman M. J .2011. *Climate governance at the crossroads: experimenting with a global response* Oxford University Press, New York.

Huitema, D., and S. Meijerink, editors. 2009. *Water policy entrepreneurs: a research companion to water transitions around the globe.* Edward Elgar Publishing, Cheltenham, UK.

Huitema, D., E. Mostert, W. Egas, S. Moellenkamp, C. Pahl-Wostl, and R. Yalcin. 2009. Adaptive water governance: assessing the institutional prescriptions of adaptive (co-) management from a governance perspective and defining a research agenda. *Ecology and Society* **14(**1): 26.

Jordan, A. and D. Huitema, 2014. Policy innovation in a changing climate: sources, patterns and effects, In: Global Environmental Change 29, 387-394

Lee, K. N. 1999. Appraising adaptive management. *Conservation Ecology* **3**:3-16.

Millo, Y., and J. Lezaun. 2006. Regulatory experiments: genetically modified crops and financial derivatives on trial. *Science and Public Policy* 33(3): 179-190.

Munaretto, S. and D. Huitema. 2012. Adaptive comanagement in the Venice Lagoon? An analysis of current water and environmental management practices and prospects for change. *Ecology and Society* **17**(2):19.

Ostrom, E. 2005. *Understanding Institutional Diversity*. Princeton University, New Haven, Connecticut, USA.

Peters, G. B. 1998. *The Experimenting Society and policy design*, p125. In **Dunn, W. (ed)** 1998. *The Experimenting Society: Essays in honor of Donald T. Campbell. Policy Studies Review Annual volume 11*. Transaction Publishers, New Brunswick, New Jersey.

Pielke Jr., R.A. 2007. *The Honest Broker: Making Sense of Science in Policy and Politics.* Cambridge University Press, Cambridge, UK.

Sabatier, P. 1987. Knowledge, policy oriented learning and policy change: An advocacy coalition framework. *Science Communication* 8(4): 649–692.

Sanderson, I. 2002. Evaluation, policy learning and evidence-based policy making. *Public Administration* **80**(1): 1 – 22.

Sanderson, I. 2009. Intelligent policy making for a complex world: pragmatism, evidence, and learning. *Political Studies* **57**(4): 699.

Tassey, G. 2014. Innovation in innovation policy management: The Experimental Technology Incentives Program and the policy experiment. *Science and Public Policy*, 41(4), pp.419–424.

Vedung, E. 1997. *Public policy and program evaluation.* Transaction Publishers, New Brunswick, USA.

Voss, J.P. and A. Simons, 2014. Instrument constituencies and the supply side of policy innovation. *Environmental Politics*, 23 (5), 735–754.

Vreugdenhil, H., J. Slinger, W. Thissen, and P. Ker Rault. 2010. Pilot projects in water management. *Ecology and Society* **15**(3): 13.

Walters, C.J. and C.S. Hollings. 1990. Large-scale management experiments and learning by doing. *Ecology* **71**: 2060–68.

Weber, M. 1968. *Economy and society: An outline of interpretative sociology*. Bedminster Press, New York, USA.

Weiss, C. 1977. Research for policy's sake: the enlightenment function of social research. *Policy Analysis* **3**(4):531–545.

APPENDIX A

Rules	Indicator	technocratic	boundary	advocacy
Boundary	*Actor Inclusiveness*	All / predominantly all expert actors	All actor types involved	All / predominantly all policy actors
	Accessibility to experiment	Invited by initiator	Requested involvement	Have organiser role/obliged
	Group members already met	Some	No	Yes
	Openness to new participants	Marginally/ some allowed	Open	Marginally/ closed
Position	*Stakeholder role*	No stakeholders	Interested parties as stakeholders	Few stakeholders
	Initiator role	Expert actors	Collaboration of actors	Policy actors
	Use of facilitator	Not used	Used	Used for select parties
Information	*Contribution to goals*	No one	All actors	Actors who are in agreement
	Lay knowledge contributed	None	Yes, to a large degree	Some, but not solely
	Scientific knowledge contributed	Majority	Some	Marginally
	Amount information received	Sufficient amount for majority of participants	Sufficient amount for all participants	Sufficient amount for minority of participants
	Opportunity for personal contact between participants	Limited	Regular	Regular for only some participants
Choice	*Authority at decision nodes*	Expert initiators	Shared power	Policy initiators
Pay-off	*How costs distributed*	Minimal buy-in	Buy-in	No buy-in
Aggregation	*How decisions made*	Experts by majority in line with scientific methods	Everyone by consensus on basis of deliberation	Policy actor by majority in reference to shared principles

Table **: The difference in rule settings for each ideal type as delineated by the institutional rules.

APPENDIX B

The changing practices of water boards

The Dutch government see climate adaptation as a response to the water issues they face and there is an urgent political need to innovate with policy solutions to meet these concerns. Change is apparent; for instance with the issues of fresh water availability and drought. Traditionally the approach to water management was to drain the country of excess water. The use of land for farming required water levels to be as low as possible so vast, efficient drainage systems were built. With the threat of climate change and the development of knowledge about the link between surface and ground water (Kuks 2002), the focus has shifted into trying to store and maintain water on the land for longer periods. These changes require governance as well as technological responses, and experiments are carried out to test these ideas. Other examples of innovative changes in Dutch water management include the fashioning of policy concepts such as multi-functional land use, which combines flood reduction and nature management; dynamic coastal management and building with nature, which uses natural processes to reduce flood risk; and water husbandry, which encourages farmers to close the water cycle and be self-sufficient with the water they have.

It is within the ambit of these responses that policy experiments were identified; however, it was not an easy task to designate a project either an experiment or some other type of project; e.g. a pilot project. The term is used quite freely in the academic literature, which is in part why this research has been undertaken, to try and understand what we really mean when we talk about experiments. The answer is we mean something specific, and actually rather rare. From an adaptive management perspective, Gunderson cites three reasons why experiments are uncommon: the natural system is not resilient to systematic testing; the social system (i.e. the political system) is inflexible; and there are significant technical challenges to designing experiments (Gunderson 1999). From the experience of hunting for them, it would appear the inflexibility (or unwillingness to fail, spend the money, and spend the time) to experiment properly would be the most common reason. Nevertheless, a sample was obtained and analysed.

Drawing from the literature, a policy experiment is expected to test causal claims, to the extent that it is able and that proponents perceive this as possible (a substantial body of literature has developed around the arguments for and against experimental evaluation to assess claims). Essentially, this test for causality is what separates experiments from other types of pilot projects, and why they are considered a superior form of evidence. However, experimenting to establish causal effects of policy changes is more straightforward in social and economic policy than environmental policy. When assessing the social system the treatments are applied to randomly chosen human actors and this can be done relatively easily, compared to the thicket of variables that need to be controlled for if randomness and control groups are attempted in the social-ecological system. Policy experiments are different from other projects by their status as pilots; however, pilot projects and policy experiments are not the same thing. The differentiation arguably lies in the extent of monitoring and evaluation. Calling a project a pilot infers that it is temporary, or first, but

most pilots are not monitored for effects, or even evaluated. A significant proportion is used for demonstration rather than testing (Sanderson 2002). It is argued here that there is room for a definition of experiment between a strict experimental design and a demonstration, which would be indicated by the presence of a monitoring and evaluation framework. The evaluation may only be of the ecosystem response, but thorough experiments will assess the social acceptance, or buy-in, of the social system as well.

Second, an experiment tests a policy innovation, in the sense of a long term alteration in policy or management practice as opposed to a mere adjustment of current practices (Duijn 2009). Policy innovations emerge from a significant concern- e.g. climate change, economic crisis- where incumbent solutions are not enough and policy makers are willing to imagine innovations that are then tested by experiment.

Third, a policy relevant innovation relates to a new policy concept or approach, indicated by a significant departure from the norm. This can come in the form of a new policy concept; such as building with nature or multi-functional land use, whereby experiments are original manifestations of the new concept in practice; or a new approach, like the shift from the state being responsible for water management to users, or the practice of storing water within the system instead of draining it.

The final three criteria were whether there was state involvement (specifically water boards), whether the experiment straddled the social ecological system by eliciting an ecosystem response, and whether it was relevant to climate adaptation. State involvement is important to declare an experiment policy relevant, and water boards were chosen because of their specific focus on water management, intention to innovate due to recently divested responsibilities, long-term focus on water management, and the fact they were involved in nearly every project assessed. Cases were looked for by searching for phrases such as: test pilots, innovation, experiment, "*proef*, *onderzoek*, pilot, on programme websites, ministry, province and water board websites, and mentioned in scoping interviews. Projects that were deemed irrelevant included product testing, concept pilots, modelling projects, and reapplications of the initial experiment.

APPENDIX C

The table notes survey responses for each experiment (Mun. = municipality; W.A. = water authority; Prov. = province; Min. = Ministry for Environment). Please note there are no experiments 12, 13, 16, and 20.

Exp #	# Init. resp.	#Know case	# Comp. resp.	Mun.	W.A.	Prov.	Min.
1	9	5	4	x	x	x	
2	9	9	6		x	x	
3	10	4	4		x		
4	8	6	6	x	x		
5	11	9	9	x	x	x	
6	9	5	4		x	x	
7	11	4	1		x		
8	4	1	1			x	
9	11	2	2	x	x		
10	11	8	4	x	x		x
11	7	7	7	x	x	x	x
14	12	8	5	x	x	x	
15	10	5	5		x		
17	7	6	5	x	x	x	
18	9	8	6	x	x		
19	10	6	6	x	x	x	
21	16	4	4		x	x	
Total (ave)	164	97 (60%)	79 (48%)	10	16	10	2