IVM Institute for Environmental Studies

Are we Learning-by-Doing Policy Experiments?

A quantitative assessment of how the institutional design of a policy experiment influences policy learning for climate adaptation.

B. McFadgen

Report R-14/42
04 December 2014

This report is released by: Professor Dave Huitema

IVM Institute for Environmental Studies

VU UNIVERSITY AMSTERDAM

This report was commissioned by: Knowledge for Climate
It was internally reviewed by: Knowledge for Climate

IVM
Institute for Environmental Studies
VU University Amsterdam
De Boelelaan 1087
1081 HV AMSTERDAM
The Netherlands
T +31-20-598 9555
F +31-20-598 9553
E info.ivm@vu.nl

Knowledge for Climate Programmabureau
Daltonlaan 400
3584 BK UTRECHT
The Netherlands
T +31 317 48 6540
F +31 6 2120 2447
E info@kennisvoorklimaat.nl

Contents

Abstract

An extensive review of the policy science and adaptive management literature reveals that little theoretical and empirical scholarship has been conducted to investigate the role of experiments in policy making and how they relate to policy learning. This is remarkable as policy experimentation is regularly, and favourably, referred to. This paper helps get a firmer grip on experiments in climate governance by presenting a novel and rigorous conceptualization and a systematic exploration of the link between experiment design in institutional terms and learning effects. Policy experiments are here defined as temporary, controlled field-trials of a policy-relevant innovation that produce evidence for subsequent policy decisions. In order to advance our understanding of experiments, a three way typology of experiments is proposed, which is informed by the policy science and science-policy interface literature: the technocratic, boundary, and advocacy (ideal) types. Institutional factors are used to demarcate the types, including their boundary, information, and choice rules as outlined by Ostrom (2005). A technocratic experiment effectively looks to separate science from power and is mainly aimed at the production of objective and generalizable scientific information. A boundary experiment embodies transparent, deliberative and inclusive processes and produces policy-relevant knowledge. In contrast, the advocacy experiment attempts to steer the process by restricting information distribution, limiting the authority of participants and protecting prevailing norms, and produces knowledge that supports a pre-determined outcome (see McFadgen and Huitema, in review).

Experiments are intended for learning, and policy learning will here be conceptualized as a measured change of understanding in the mind of an individual who is part of a policy community. In this paper we aim to gauge learning effects and we do so by developing hypotheses on how the three types of experiments might be connected to various types of learning. We borrow from new advances in the literature aimed at isolating certain factors that affect learning, such as the level of diversity of actors; the openness and sustained exchange of information; the degree to which control over the process/joint fact-finding is shared; the presence of facilitation; and the existence of consensus decision making (Mostert et al. 2007; Gerlak and Heikkila 2011; Muro and Jeffrey 2012; Leach et al. 2013; Baird et al. 2014). The key and novel hypothesis advanced here is that technocratic experiments produce high cognitive learning, no normative learning, and some relational learning, while evidence emanating from this type of experiment will be perceived in the policy network as credible. We hypothesize that boundary experiments produce some cognitive learning, high normative learning, and high relational learning, and that the evidence produced will be perceived as salient and legitimate. We finally submit that advocacy types will produce low cognitive, some normative, and low relational learning, and evidence that is salient.

An empirical study investigated the use of policy experiments in the Dutch climate adaptation policy field and found 18 experiment cases out of 174 innovative initiatives, indicating that experiments are actually quite a rare phenomenon in this field. The experiments tested policy concepts tackling coastal, flooding, drought, and fresh water availability issues and were dated from 1997-2012. An assessment based on 20 institutional indicators revealed that four experiments can be classified as technocratic, seven as boundary type, and seven as advocacy type. This is far fewer technocratic and far more boundary types than expected, indicating that experiments are used to bring different actor types into the policy process, and they produce more policy relevant knowledge than strictly scientific knowledge. It also found that experiments vary in terms of their social, bureaucratic, and political dynamics; including how open they were to outsiders, the legal barriers they faced, and how controversial their intentions were.

An analysis of resultant learning shows that ideal types produce significantly different learning effects and that the hypotheses are largely met. However, in all experiments there was very little change in the values and norms of participants (compare Huitema et al., 2009), which raises questions as to why a deliberative setting could not encourage this form of normative learning. The technocratic type produced most cognitive learning but conversely a boundary experiment produced the least cognitive learning, illuminating a possible trade-off between knowledge acquisition and trust building. As expected, advocacy experiments produced the least relational and normative learning, and were significantly less likely to be controversial, indicating the policy actors may use this design for structured, uncontentious policy issues.

1 Introduction

Climate pressures are causing policy makers in environmental governance sectors to search for new and improved approaches to managing climate change. In particular, adaptation is gaining traction as the new policy focus alongside mitigation, and like for other environmental concerns, learning is encouraged as a success criterion (Baird *et al.* 2014).

Learning is a popular research subject, with theoretical underpinnings in the academic fields of sociology, education, commerce, and environment; thus scholars call for clarity on its use (Reed *et al.* 2010; Rodela 2011; Leach *et al.* 2013). For the purposes of this research, policy learning is defined as: "relatively enduring alterations of thought or behavioural intentions that result from experience and that are concerned with the attainment (or revision) of public policy" (Sabatier 1988) and one enabler of this sort of learning is experimentation. Experimentation brings new and reliable knowledge into the policy process by testing innovative policy options on a short term basis without committing to a specific course of action, thereby improving understanding of alternative approaches. It has the potential to save money and save face, by detecting unexpected consequences before policy commitments are made (Voß & Bourneman 2011; Tassey 2014). How experimentation might enable learning is a research question not often tackled in environmental governance literature, and it spawns further questions, such as: what might this learning look like; and are there factors that are more important than others in generating learning outcomes?

Against this backdrop, the aim of this article is to investigate whether there is a relationship between the way experiments are designed and the kinds of learning effects they generate. The research builds on recent studies of learning in environmental governance, such as Muro & Jeffrey, 2012; Leach *et al.* 2013; and Baird *et al.* 2014, where learning is explicitly conceptualised and measured in specifically chosen learning situations. However, this research departs from the studies listed above due to the observed setting. Baird *et al.* (2014) examines learning in a participatory decision making process, Leach *et al.* (2013) in collaborative partnerships, and Muro and Jeffrey (2012) in participatory work groups. In contrast, this study compares learning outcomes in three alternative settings, thereby allowing us to ask, among other questions, how much better are participatory institutional settings for learning than other institutional designs?

2 Theory and framework

2.1 Experiments

The necessity for governments to get their policy decisions right first time is greater than ever, with climate change and other urgent environmental issues pushing hard on the policy agenda. Science and expertise is called upon to enlighten policy actors on what risks we face and what consequences their decisions will have and whole industries have been built around this service to government.

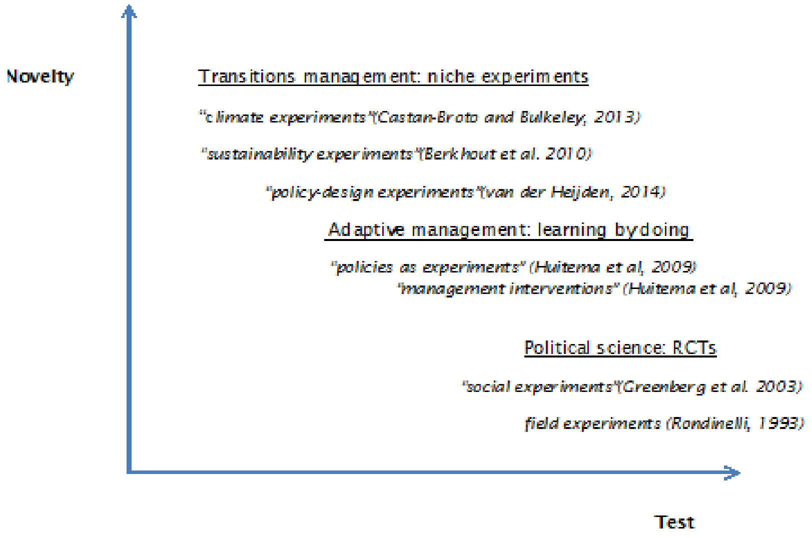
One method governments can employ to aid their decision making role is by commissioning experiments. Policy experimentation has an extensive political and academic history, with analyses conducted since the 1960s when the idea of the Big Society took shape. DT Campbell was one of the first to kick against what he saw as policy decisions being taken without the risk of criticism or failure and he advocated the use of policy evaluation; the most rigorous form being the random controlled experiment (Campbell 1998). 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. Scientific experts designed the experiments and conducted them on citizens to generate evidence of what policies work to minimise social problems. Criticisms of the use of experiments essentially revolved around ethical issues of such experiments and the belief that the complexities of the social world could be understood and managed through such a limiting prism (e.g. Fischer 1995; Sanderson 2002), after all the real world cannot confine itself to laboratory type conditions (e.g. no true generalizability or counterfactual). Thus, any results from social policy experiments could not be considered reliable and the method lost support.

In environmental policy, another discussion about experimentation took place under the adaptive management approach. Here, experimentation was also supposed to provide reliable evidence about policy interventions that worked, but the focus was on policy actors and experts working together in experiments and later extended to incorporating a broader range of non-state actors in the adaptive co-management approach (Lee 1999; Armitage *et al.* 2008; Huitema *et al.* 2009).

A third, broader understanding of experiments has recently emerged that focuses less on generating reliable evidence and more on seeking innovative policy approaches. Climate experiments and transitions management experiments aim to implement novel solutions within spaces either protected by market forces or outside the ordinary policy process (Berkhout *et al.* 2010; Castán Broto & Bulkeley 2013). A broad range of actors are brought into the process and seek to change policy actions through the implementation of shadow networks (Meijerink & Huitema 2007).

These three understandings of experimentation emphasise varying degrees of its testing and novelty function, but have in common the positive characteristic of being flexible, in that they are not implemented on a full scale and are reversible (Tassey 2014). A useful definition that captures these varying characteristics of policy experimentation is:

“a temporary, controlled field-trial of a policy-relevant innovation that produces evidence for subsequent policy decisions”.


Novelty
Transitions management: niche experiments
"climate experiments"(Castan-Broto and Bulkeley, 2013)
"sustainability experiments"(Berkhout et al. 2010)
"policy-design experiments"(van der Heijden, 2014)
Adaptive management: learning by doing
"policies as experiments" (Huitema et al, 2009)
"management interventions" (Huitema et al, 2009)
Political science: RCTs
"social experiments"(Greenberg et al. 2003)
field experiments (Rondinelli, 1993)
Test

Graph 1 Comparison of experiment conceptualisations against the axes of testing and novelty, to show concept diversity. The test axis refers to the extent an experiment type emphasises the role of monitoring and control; and novelty refers to the extent a concept emphasises innovation. These categorizations are rough and their placement in the two-dimensional graph should be seen as approximate (taken from Crona & Parker, 2012).

This research concentrates on experiments that have a testing function and relate to policy by producing evidence for public policy decisions. The relevance to policy is important because of this report's focus on the positioning of experiments as institutional arrangements at the science policy interface.

Use in policy development

One recognisable function of an experiment is its intention to provide evidence of how an alternative policy approach will work. In environmental governance, understanding how both the social as well as ecological system will behave is crucial, and by implementing a project on a temporary basis and monitoring it for effects we enhance our understanding of relevant social-ecological system dynamics, as well as the policy approaches to see if they deliver desired outcomes (Steffen 2009). This flexibility allows us to govern better under uncertainty (Anderies and Janssen 2013). In the literature uncertainty has been disaggregated into three types: incomplete knowledge, unpredictability of the system, and ambiguity of the solution (van Hoek *et al.* 2013) and by picking up data on the intervention's actual effects, experimentation can arguably address all three forms.

Experiments might not only lay evidence at the feet of policy makers, they may also provide the opportunity for the divergent worlds of policy actor and expert to mix and discover new ways to manage the environment. As discussed above experiments vary in to what extent they involve other actors. If used as a platform to invite other actors into the policy process then experiments can enable creative inquiry and discovery, and the inclusion of public judgement in policy decision making (Dryzek 1987; Caspary 2002). Experiments offer the opportunity to involve other knowledge types and other points of view; involving stakeholders that might otherwise oppose (Lee, 1993; Petersen *et al.* 2011).

Finally, there is a political aspect that, despite attempts to make experiments appear neutral, cannot be ignored. On behalf of society the state determines what issues sit on the policy agenda and this status is often a driver of the choice to commission an experiment. Experiments are conducted to gather evidence but that evidence may be used to support a predetermined decision and soften objections, since a temporary change that is reversible provides a sense of security and renders actors less resistant to policy change (Vedung 1997; Tassey 2014). Experiments can be used to push an idea or maintain an idea on the policy agenda, or they could be used as a tool to delay making final decisions (Greenberg *et al.* 2003).

2.2 Learning

In the policy sciences and environmental governance literatures, learning has emerged as a mainstay for improving decision making (Cundhill and Rodela 2012; Leach *et al.* 2013). It can lead to preferred states, like increased adaptive capacity (Lebel *et al.* 2010) or it can be a normative outcome in itself; i.e. embodying a change in thought in a policy community from new knowledge (Huitema *et al.* 2010). Drawing on Lee (1999) two associations come to mind when considering learning in relation to experiments. First, an experiment's monitoring and testing mechanisms aim to produce acute and accelerated knowledge production, improving understanding of the social and ecological systems response to an intervention. Second, owing to its novelty an experiment may be seen as a unique configuration of issues and participants, with the potential to create new actor networks that engage non-state actors in a policy decision making process (Leach *et al.* 2013).

It is important to the advancement of learning theory that researchers are explicit as to what learning they are focusing on (Leach *et al.* 2013). As Crona and Parker (2011) lament, there is little consensus on learning definitions nor on factors that foster learning. However, in recent years scholars of environmental governance have been making a specific attempt to categorise and measure learning among individuals, often in participatory settings (for a discussion on the difference between this conceptualisation and broader learning settings, such as network or systems centric perspective, see Rodela 2012). Three studies of learning in particular have sought to define learning and better understand triggers that encourage it, namely Muro and Jeffrey (2012); Leach *et al.* (2013); and Baird *et al.* (2014).

3 Conceptual framework

With these conceptual understandings in mind, the main research question investigated by this article is:

To what extent can policy learning be explained by an experiment's institutional design?

In order to answer the question, an hypothesis is constructed that proposes learning is dependent on the type of design an experiment has. The dependent variables are three different learning effects important to policy development, measured at the level of participant in an experiment. The three effects are: cognitive, normative, and relational learning (Huitema *et al*. 2010; Haug *et al*. 2011; Baird *et al*. 2014b). Cognitive learning refers to an individual's knowledge acquisition and increased complexity of understanding. New information serves both advocacy and enlightenment functions for policy making (Grin and Loeber 2007) and cognitive learning represents the uptake of new knowledge that reduces uncertainty of an experiment's effects on the social-ecological system. Normative learning refers to a change in an individual's values, goals, or belief systems, with value in building a common interest among participants (Leach *et al*. 2013). Relational learning is understood as an improvement in understanding of others' mindsets and an increase in trust and cooperation, with trust being a vital component in governing social-ecological systems (Poteete, Janssen and Ostrom 2010).

Table 1 Policy learning effects, their definitions, and added value to the policy process.

Typology of policy learning	Definition	Value to policy process:
Cognitive learning	Acquisition of new knowledge and restructuring of existing knowledge	Reduces uncertainty. Information provides advocacy and enlightenment functions.
Normative learning	Change in norms, values, goals.	Synthesis of priorities between individuals about the policy issue that leads to a common interest or goal built within the group.
Relational learning	Enhanced trust, improved understanding of mind-sets of others.	Importance of trust in governing social-ecological systems.

The hypothesis's independent variables are experiment types that draw on institutional rules provided by Ostrom (2005); i.e. those of an action situation that determine who is involved (boundary rules), what authority they have (choice rules), what information they share (information rules), and what positions they hold (position rules, pay-off rules). These rules can be set at different modes, which collectively delineate three types of design based on concepts drawn from the science-policy interface and policy sciences (Dryzek 1987; Pielke 2007). The ideal types are explained below and table 2 sets out more detail with the different rule settings for each of the types. The hypothesis is shown in figures 1 and 2, followed by explanation of the theory underpinning it.

Technocratic type

A technocratic experiment resembles the technical-rational model of policy making, where there is a separation of power between the experts who provide knowledge and the policy actors who make decisions based on that knowledge (Owens *et al*. 2004). Scientists play a vital but objective and disconnected role in politics as ‘pure scientists’ (Pielke 2007). Expert actors are the initiators and sole participants of a technocratic experiment and maintain control over its design, monitoring, and evaluation. Although policy actors are commissioners of the research, they are absent or supporting the experts and do not have decision authority. Scientific knowledge is the sole type of information generated by the experiment and there is no constructing or reflection on policy goals as they are already established by policy actors in advance.

Advocacy type

In contrast, an advocacy experiment is initiated and controlled by policy actors and is used to transfer select information and bring particular actors into the policy process in a form of ‘stealth advocacy’ (Pielke 2007). Although appearing neutral, the experiment supports particular outcomes as participants must be invited and outsiders are barred from gaining access (Owens *et al*. 2004). Participating actor types may be diverse but they are carefully selected to provide specialist knowledge and are mostly from the same actor network. They have little influence over the design, monitoring and evaluation procedures, reinforcing existing structures of power. Within the group, only a few participants discuss and shape goals and share information through the use of a facilitator so prevailing norms are protected. Information distribution is generally suppressed within the group as well as from outsiders.

Boundary type

A boundary experiment adheres to the principles of democracy and legitimacy by opening itself up to any actor- state or non-state- that has a desire to be involved in policy making. It is initiated by a collaboration of actors and generates diverse knowledge types, including ordinary and practical knowledge from non-experts, making it more responsive to policy needs. 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. Reflexivity is high in a boundary experiment because participants contribute to the discussion on appropriate goals and whether the experiment adheres to acceptable societal aims. Deliberative practices are encouraged with transparent information transmission, open dialogue, and regular communication among participants.

Table 2 The difference in rule settings for each ideal type.

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	Few actors
	Lay knowledge contributed	None	Yes, often solely	Some, but not solely
	Scientific knowledge contributed	Majority	Some	Marginally
	Amount information received	Majority found sufficient	Everyone found sufficient	Minority found sufficient
	Opportunity for personal contact	Sometimes	Often	Rarely
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	Everyone by consensus	Policy actor by majority

3.1 Hypotheses

With these variables in mind, the hypothesis proposes that an experiment ideal type has influence over subsequent learning effects.

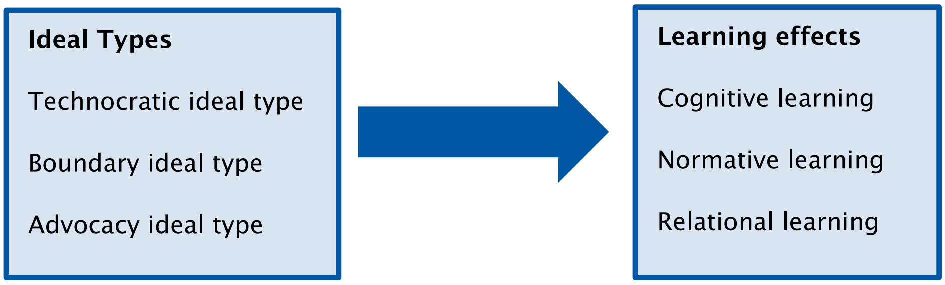


Figure 1 Experiment ideal types have an effect on the learning outcomes of a policy experiment.

This hypothesis is informed by design factors uplifted from the learning literature, in particular the field of social learning (figure 2). Several factors that are said to contribute to learning, with the relevant ones in effect being tested here. They include a diversity of actors to share knowledge and views; open and sustained exchange of information; control over the process/joint fact-finding; facilitation; and consensus decision making (Mostert *et al*. 2007; Gerlak and Heikkila 2011; Muro and Jeffrey 2012; Leach *et al*. 2013; Baird *et al*. 2014). Control factors would be those beyond the scope of institutional design choices; for instance, a charismatic leader, demographics, the participants' willingness to learn, and media attention.

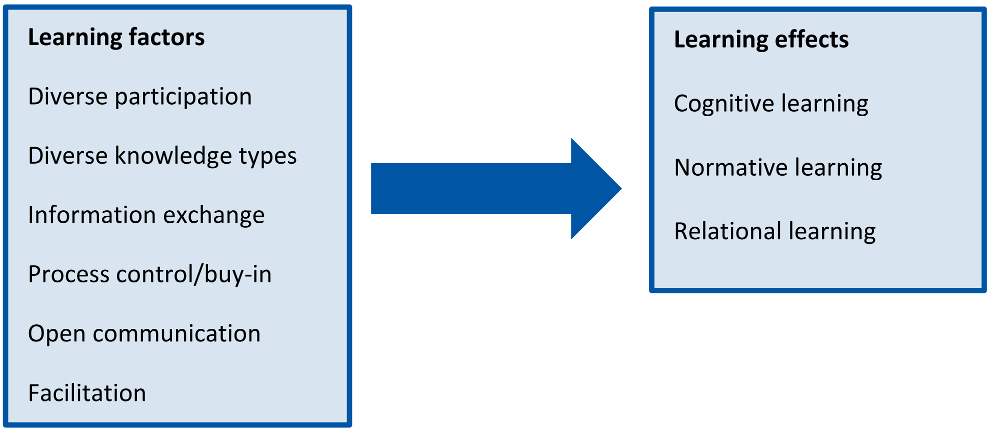


Figure 2 *Specific factors from the literature that are varied in the ideal types to see if they create different learning effects.*

With its low participant diversity and myopic focus on generating generalizable scientific evidence a technocratic design will generate high levels of cognitive learning but low normative learning. The relatively open exchange of information and shared authority allows for some relational learning. An advocacy design allows for diversity in actor types but their influence is very low and information is not openly shared, generating some cognitive and normative learning but low levels of relational learning among the group. A boundary experiment has high levels of normative and relational learning because of the focus on diverse actors openly contributing knowledge and views while sharing authority, financial buy-in from the participants, and the use of a facilitator. However, cognitive learning is somewhat affected by the drive to produce reflexive information over instrumental, scientific information.

With these theoretical underpinnings, the test hypothesis can be broken into sub-hypotheses:

> *A technocratic experiment produces the highest levels of cognitive learning, lowest normative learning, and some relational learning within the circle of participants in the experiment.*
>
> *An advocacy experiment produces some cognitive learning, some normative learning, and the lowest levels of relational learning.*
>
> *A boundary experiment produces some cognitive learning and the highest levels of normative and relational learning within the circle of participants in the experiment.*

In order to test the validity of these hypotheses, 18 experiments were analysed, with the methods and results presented in the following sections.

4 Methods

The search for policy experiments was conducted throughout various institutions of the Netherlands, with a focus on the country's water authorities ("water boards"). 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 water boards sit at the regional level between local and provincial government and form a fourth institutional level that is, until recently, somewhat unique to the Netherlands. The main responsibilities of the water boards are the maintenance of dikes and dams, water quantity and water quality. For this report the first two tasks are seen as most relevant to climate adaptation.

4.1 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 report was written, 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, spend the time) to experiment properly would be the most common reason. Nevertheless, a sample was obtained and analysed, and the criteria used to select experiments is outlined in the next section.

4.2 Case selection

Six criteria were used to identify experiments: 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 elaborated on in turn.

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. Moreover, regular state of the environment assessments mean that the ecological system is being monitored anyway and provides evidence for the effectiveness of environmental policies so further controlled evaluations could be seen as superfluous. 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 are 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.

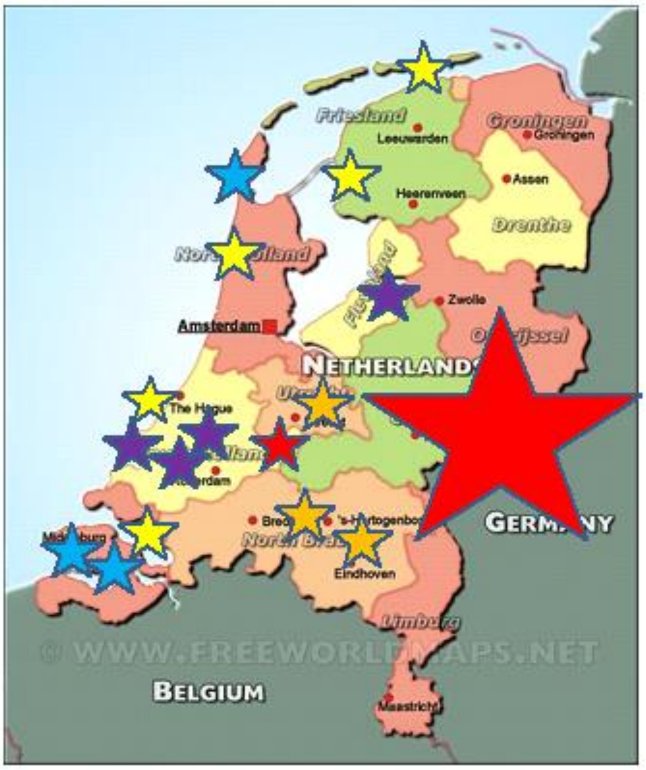
174 innovative pilot cases were identified according to the criteria and 18 cases were selected as meeting all six criteria, the most uncommon criteria being a monitoring

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 these stringent criteria.

Experiment cases

The 18 cases of experimentation in Dutch climate adaptation test policy innovations in coastal and inland defence, flooding, and drought related issues. They date between 1997 – 2012 and nearly half are ongoing. Ongoing cases are included if they have passed at least one evaluation phase. The names of the experiments are not given to honour confidentiality; however, map 1 below illustrates the location of the experiments in the Netherlands and the colour of the stars shows what water issues they relate to. Yellow stars indicate coastal experiments, blue stars water availability experiments, purple is multi-functional land use experiments, orange stars delineate water variability experiments, and the red stars show dike management experiments, with the red star delineating an experiment that is being conducted throughout the country.

The following sections detail data collection and survey protocols that were followed.



Map 1 Map of the Netherlands showing locations of each experiment case. Yellow stars show location of 5 coastal management experiments; purple stars show 4 water storage experiments; blue stars show three freshwater experiments; orange stars show three water variability experiments; and 2 red stars show dike management experiments. The giant red star denotes an experiment conducted in several regions of the country.

5 Data collection

Participants of the experiments were identified during interviews with project leaders and analyses of project reports. Participant numbers varied across cases, with the smallest having eight participants and the highest with 40. Experiment participants were sent via email an online survey with 64 closed questions, which asked about their role in the experiment, their opinions on design aspects, and questions to gauge their learning experiences. Three reminder emails were sent at weekly intervals. A total of 265 survey emails were sent out and 170 were completed, giving a 64% response rate. Each case either had a minimum of 6 responses or responses from over half of known participants.

Survey design

The respondents were asked a mixture of factual and attitudinal questions to gauge the setting of the institutional rules; including questions about the role of the respondent, the design of the experiment, the other participants, power structures, and financing of the project. Control questions were also asked; including how controversial the project was, competency of the initiator, legal barriers, and media attention.

Learning assessment

Table 3 states the learning variables measured, the survey questions/statements used to measure the learning variables and the correlation statistic for each set. All sets of questions were significantly correlated ($0.19 < r < 0.48$) so they were grouped together to reduce the number of variables from 11 to 6. The analysis treats the scales as continuous. The questions follow closely the conceptual framework in table 1 above.

Table 3 Questions asked to gauge learning and correlations cores to justify grouping.

Cognitive learning 1- participant's knowledge acquisition.
Scale: -2= strongly disagree; -1= disagree; 0=neutral; 1= agree; 2= strongly agree.
I gained new factual information from the experiment.
Through participating in the experiment, I gained better insight into the manner in which environmental problems can be solved.
Correlation= 0.28 (sig.)
Cognitive learning 2- participant's improved complexity of understanding.
Scale: -2= Not at all; -1= Slightly; 0=A certain extent; 1= Strongly; 2= Very great extent.
By participating in the experiment, did you improve your personal knowledge of the natural system in question?
To what extent have the results of the experiment surprised you and forced you to amend your initial expectations about the outcomes of the intervention? Correlation= 0.35 (sig.)
Normative learning 1- participant/group change in priorities.
Scale: -2= strongly disagree; -1= disagree; 0=neutral; 1= agree; 2= strongly agree.
Participating in the experiment has changed the importance I attach to environmental issues.
Participating in the experiment has changed the importance others attach to environmental issues.
Correlation= 0.19 (sig.)
Normative learning 2- group formed common interest.
Scale: -2= strongly disagree; -1= disagree; 0=neutral; 1= agree; 2= strongly agree.
The experiment ensured that participants discovered a common goal.
Relational learning 1- participant better understood others' mind-sets
Scale: -2= strongly disagree; -1= disagree; 0=neutral; 1= agree; 2= strongly agree.
As a result of the experiment I got a better understanding of the mind-sets of other participants.
Even though differences remain, I have developed a bond with my opponents by participating in the experiment.
Correlation= 0.34 (sig.)
Relational learning 2- participant/group increased trust
Scale: -2= strongly disagree; -1= disagree; 0=neutral; 1= agree; 2= strongly agree.
As a result of the experiment a mutual trust has grown between participants. *I would participate again in an experiment with these participants.*
Correlation= 0.48 (sig.)

6 Results

This section first evaluates whether learning took place in the sample, then it assesses the ideal type for each experiment before comparing the ideal type characteristics found in the sample with those defined conceptually earlier in the paper. The next step is the statistical analysis, which establishes to what extent the ideal types (*ipso facto* the varied institutional designs) explain the differences- if any- in learning scores between the experiments.

First, to what extent did learning take place in the sample? Table 4 sets out the mean and standard deviation for each learning variable. These results show that on average, participants in the experiments learned. The gaining of knowledge and improvement in trust scored highest, with the changing of priorities clearly scoring lowest. In line with previous studies, normative learning proved difficult to achieve (Leach *et al.* 2013; Baird *et al.* 2014).

Table 4 Learning scores for the six types. Scale used: -2= strongly disagree; -1= disagree; 0=neutral; 1= agree; 2= strongly agree[1]*.*

	N	Minimum	Maximum	Mean	SD
Cognitive learning- knowledge acquisition (CL1)	153	-1.5	2	**0.7**	0.65
Cognitive learning- improve understanding (CL2)	153	-2	2	**0.0**	0.7
Normative learning- priorities (NL1)	152	-2	1.5	**-0.1**	0.64
Normative learning- convergence (NL2)	152	-2	2	**0.6**	0.71
Relational learning- understand other's mind-sets (RL1)	152	-1.5	2	**0.3**	0.59
Relational learning- trust (RL2)	152	-2	2	**0.8**	0.75

Next, 21 indicators were identified to assess the cases based on an experiment's institutional design. As explained in the concept section, the indicators have three rule settings, one for each ideal type. Table 1 above sets out how the indicators conceptually delineate ideal types, and Appendix A lists a more detailed breakdown of indicators by rule category and their setting for each ideal type.

In order to determine what type an experiment is, the experiments were assessed and given a score for each indicator. The scores correlate with a particular setting, and the experiment is labelled the ideal type that its scores best correspond to. For example, Experiment 2 had three scores for technocratic, 14 scores for boundary, and 4 scores for advocacy type; thus it is a boundary experiment. The assessment concludes that for the 18 cases, four experiments meet the technocratic definition, seven are boundary experiments, and seven are advocacy experiments (table 5 below).

Table 5: Number of cases that fall into each ideal type.

[1] Cognitive learning 2 used a different scale, -2= Not at all; -1= Slightly; 0=A certain extent; 1= Strongly; 2= Very great extent.

Table 5: Number of cases that fall into each ideal type.

Ideal type	Number of experiments that meet the type
Technocratic	4
Boundary	7
Advocacy	7

For a visual representation of the categorization see figure 3, which shows each case plotted in a ternary plot. Red crosses are cases with mostly boundary characteristics; green crosses are technocratic cases; and blue are cases with mostly advocacy characteristics. The placements are calculated by plotting in the triangle what scores for each type the experiments gained. Figure 3 shows that it is uncommon for cases to fall absolutely into one type. It is more common for cases to display characteristics of two types, hence some of the cases are hovering towards the middle of the figure. The figure also shows that experiments tend not to have technocratic scores compared to the other types. Despite these findings, the plot shows that cases do cluster based on the indicator assessment and that three types can be identified.

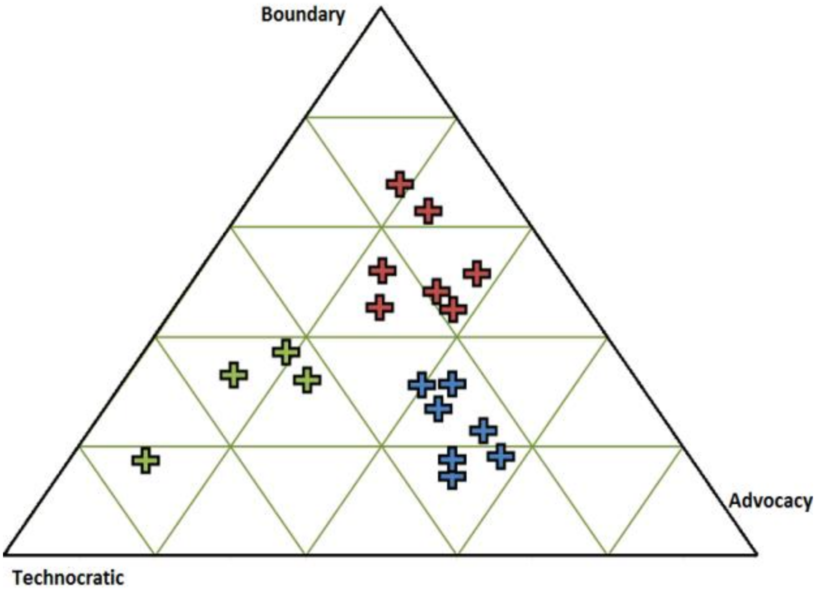


Figure 3 ternary plot to illustrate how the cases fall into one of three ideal types.

Table 6 shows how the experiments scored in relation to the ideal types. It sets out the experiment scores for each indicator and shows which ideal type was most common for the different indicators.

The first set of indicators are for the position rules, which prescribe what sorts of positions participants can take in the experiment. The sample showed that for the position rules the most common ideal type is the advocacy type. The technocratic type is least common. For example, only one experiment had experts initiate the experiment, with most being initiated by policy actors. The next set of rules are the authority rules, where the three different decision nodes are about design choices, monitoring and implementation, and evaluation. Here the technocratic type is most common, because experts are generally taking decisions at these nodes (often in conjunction with another actor type). Advocacy experiments also scored high, when policy actors take decisions. The least common is the boundary type, where three or more actor types make decisions together on specific nodes. This indicates that joint decision making and process control are rare.

Information rules prescribe what sorts of information is generated and the distribution of that information. Here the boundary type is most common, as actors generally felt they received sufficient information and that it was openly shared amongst all participants, which is a surprise since it was expected that answers for this variable would be more diverse. Reflexivity, the extent goals are discussed and shaped by participants, scored highest for advocacy types because generally only a select group of participants shared their views on goals. Aggregation rules describe how decisions are made- whether they are unilateral or shared, for generic decisions about amending or ending the project. Again, advocacy experiments score best, which means most experiments had unilateral decision structure as opposed to shared or consensual.

Payoff rules prescribe the extent there is buy-in from participants or whether the initiator pays costs, with technocratic experiments, those where expert actors are paying for the experiment, score best. Finally, boundary rules determine who is let into the experiment, and thereby involved in the policy process, and who is left out. The results show that experiments that are predominantly expert are most uncommon, with most cases involving several actor types. However, around half the participants knew each other, with slightly more actor groups being produced than being reinforced. The cases were also more likely to allow new participants to join during the experiment than not. However, more cases had closed access, meaning participants had to be invited into the project, with only 6 cases having participants that requested entry.

Table 6 How the ideal types score with each set of indicators derived from institutional rules.

Rule type	Identifier	Indicator	TIT score	BIT score	AIT score
Position	1A	Stakeholder position	2	8	8
	1B	Initiator type	1	10	7
	1C	Use of facilitator	5	3	10
Authority	2A	Design authority	8	2	6
	2B	Monitoring and evaluation authority	7	3	6
	2C	Evaluation authority	8	1	9
Information	3A	Reflexive discussion on goals	2	6	10
	3B	Extent of lay knowledge	3	7	8
	3C	Extent of scientific knowledge	8	5	5
	3D	Sufficient information sharing	4	12	2
	3E	Information openly shared	6	11	1
	3F	Extent personal contact	6	11	1
	3G	Outsiders informed	5	11	2
Aggregation	4A	Amend power	3	4	11
	4B	End power	3	5	10
Payoff	5A	Cost distribution	8	6	4
Boundary	6A	Inclusiveness	3	7	8
	6B	New actor group	8	6	4
	6C	Describe others	5	8	5
	6D	Openness to others	7	8	3
	6E	Access	5	6	7

The analysis above highlights what indicators of the conceptual ideal types (as described in table 2) are most common empirically. Most surprising is the diversity of actors and information generated, and how open the participants found the information distribution process. More stakeholders were involved than expected, and it is interesting to see how collaborative the initiators were. With such diversity it is surprising that so few experiments had consistent facilitation, or that few new actor networks emerged. However, the authority distribution is very low for non-state and non-expert actors, which reveals that although diverse actor groups are involved their influence is fairly narrow.

The assessment above described the set of experiments under analysis and how their design characteristics differ within the sample. In order to test the hypotheses, the next step is to compare the learning scores of the different types.

Ideal types and learning

The null hypothesis is that there is no relationship between the learning effects and ideal types. The alternative hypotheses are that there is a relationship, and that cognitive learning is significantly higher in technocratic experiments compared to the other types, and normative and relational learning is significantly higher in boundary experiments compared to the other types. As a refresher, the hypotheses to be tested are shown here in Table 7:

Table 7 Expected levels of learning for each ideal type.

	Cognitive learning	Normative learning	Relational learning
H1: Technocratic type	Highest	Lowest	Middle
H2: Boundary type	Middle	Highest	Highest
H3: Advocacy type	Lowest	Middle	Lowest

Table 8 shows the learning scores for each ideal type, with the learning classification beneath to check against the hypotheses. The table shows that technocratic experiments generate levels of cognitive learning that meet expectations. There is some normative learning in TITs, which is contrary to H1, and relational learning partially meets expectations. H2 predicted highest relational and normative learning scores for boundary experiments, which is confirmed by results; however, the low cognitive learning scores are a surprise. Finally, H3 expected the advocacy experiments to have the least learning, and their low normative learning meets expectations and relational learning partially. However, advocacy experiments produce quite a bit more cognitive learning than anticipated. Table 9 illustrates where the hypotheses are correct and where they are incorrect.

Table 8 Means of learning scores for each ideal type.

	CL1	CL2	NL1	NL2	RL1	RL2
Technocratic type	0.95	0.65	-0,1	0.6	0.15	0.75
	Highest	*Highest*	*Middle*	*Middle*	*Lowest*	*Middle*
Boundary type	0.6	-0,2	0.0	0.8	0.5	0.9
	Lowest	*Lowest*	*Highest*	*Highest*	*Highest*	*Highest*
Advocacy type	0.7	-0,1	-0,2	0.5	0.2	0.7
	Middle	*Middle*	*Lowest*	*Lowest*	*Middle*	*Lowest*

Table 9: Whether the learning means confirm or do not confirm the research hypotheses.

	Cognitive learning	Normative learning	Relational learning
H1: Technocratic type	✓	X	X/ ✓
H2: Boundary type	X	✓	✓
H3: Advocacy type	X	X	X / ✓

The results above assume that the model is robust. In order to test whether the different learning scores are in fact a result of the delineation of types, a one-way ANOVA, robust ANOVA Welch Test, and post-hoc tests (Tukey and Games Howell) were conducted. Table 10 sets out the results (significant results in bold).

The results show that the scores for normative learning 1 (priorities) and 2 (common interest) and relational learning 2 (trust) cannot be explained by the different ideal types and we cannot statistically reject the null hypothesis.

However, for cognitive learning 1 and 2 and relational learning 1, there was a statistically significant difference between means ($p < .05$) so the null hypothesis is rejected and the alternative hypothesis accepted for these learning variables. Post hoc tests show there will be a significant increase in knowledge acquisition (CL1) if the experiment is a technocratic experiment and not a boundary experiment. Post hoc tests for CL2 state significantly that if the experiment is a technocratic type then participants will improve their existing understanding of the facts. Finally, results show that the differences in understanding mind-sets (RL1) are significant between the types, but whether one is superior is not significant.

Table 10 Results of ANOVA analysis and post-hoc tests for each of the learning effects.

CL1- Knowledge acquisition	***Cognitive learning 1 was statistically significantly different between different ideal types, F(2,150) = 3.325, p < .0039****. There was homogeneity of variances, as assessed by Levene's Test of Homogeneity of Variance (p = .122). Data is presented as mean ± standard deviation. Cognitive learning 1 improves from the boundary (0.6 ± 0.7), to advocacy (0.7 ± 0.6), to technocratic (1.0 ± 0.5) types, in that order.* ***Tukey post-hoc analysis revealed that the mean increase from boundary to technocratic (0.37, 95% CI [0.03, 0.7]) was statistically significant (p = .029)*** *but no other group differences were statistically significant.*
CL2- Increased complexity in understanding	***Cognitive learning 2 was statistically significantly different between different ideal types,*** *F(2,150)* ***= 19.018,*** *p* ***< .0005.*** *There was homogeneity of variances, as assessed by Levene's Test of Homogeneity of Variance (p = .353). Data is presented as mean ± standard deviation. Cognitive learning 2 improves from the boundary (-0.2 ± 0.7), to advocacy (-0.1 ± 0.6), to technocratic (0.6 ± 0.5) types, in that order.* ***Tukey post-hoc analysis revealed that the mean increase from boundary to technocratic (0.85, 95% CI [0.5, 1.18]) was statistically significant*** *(p* ***= .0005), and the mean increase from advocacy to technocratic (0.72, 95% CI [0.36 – 1.07] was statistically significant (p= .0005),*** *but no other group differences were statistically significant.*
NL1- Change in priorities	*Normative learning 1 increased from the advocacy (-0.20 ± 0.6), to technocratic (-0.09 ± 0.7) to boundary (0.01 ± 0.7) ideal types, in that order, but the differences between the ideal types was not statistically significant, F(2,0.704) = 1.748, p = .178.*
NL2- Forming of common interest	*Normative learning 2 increased from the advocacy (0.49 ± 0.8), to technocratic (0.57± 0.6) to boundary (0.77 ± 0.7) ideal types, in that order, but the differences between the ideal types was not statistically significant, Welch's F(2, 76.9) = 2.204, p = .117.*
RL1- Understanding of mind-sets	***Relational learning 1 was statistically significantly different between different ideal types, Welch's*** *F****(2, 65) = 4.058,*** *p* ***< .022.*** *The assumption of homogeneity of variances was violated, as assessed by Levene's Test of Homogeneity of Variance (p = .048). Data is presented as mean ± standard deviation. Relational learning 1 improves from the technocratic (0.1 ± 0.7), to advocacy (0.2 ± 0.8), to boundary (0.5± 0.6) types, in that order. Games-Howell post-hoc analysis revealed that none of the changes were statistically significant (although the increase from technocratic to boundary type (0.36, 95% CI (-0.012 – 0.731)) was almost statistically significant (p = .067), as well as the increase from advocacy to boundary type (0.3, 95% CI (-0.023-0.616), p = .074).*
RL2- Increase in trust	*Relational learning 2 increased from the advocacy 0.72 ± 0.8), to technocratic (0.77± 0.6) to boundary (0.91 ± 0.5) ideal types, in that order, but the differences between the ideal types was not statistically significant, Welch's F(2, 64.2) = 1.659, p = .198.*

7 Discussion

The purpose of this article is to explore the relationship between policy experiments and policy learning using an original conceptual model that connects three different institutional design settings and three different learning effects. Results of an online survey from 170 participants in 18 cases revealed that experiments can be categorised as either technocratic, boundary, or advocacy types; with advocacy and boundary types being most common. The hypotheses were partially met, with some unexpected results that will now be reflected on with discussion regarding implications for theory.

The first point relates to the finding that experiments produce learning effects, which is a welcome one, if not wholly unsurprising. Knowledge acquisition and increased complexity in understanding meets with the assumption in adaptive management that experiments provide knowledge to reduce uncertainties (Cundill & Rodela 2012). The scores for trust building and better understanding others' mind-sets also chimes with theories that experiments can also provide platforms for new actor groups to solve policy problems together (K. Lee 1999). However, the scores for a change in the importance a participant or the group attaches to environmental issues were woefully low, despite boundary experiments containing a diverse set of actors with shared authority, exposure to a range of different knowledge types and discussions on goals, with open communication and sufficient knowledge exchange.

Caution can be raised as to the methods of measuring what is in effect deliberative processes using quantitative methods (Haug *et al.* 2011), but this does not completely undermine the clear result. Moreover, it is much in line with previous learning studies. Haug *et al.* (2011) found no evidence of normative learning from conducting a policy game exercise. Likewise, Munaretto & Huitema (2012) found no evidence of normative learning from experiments conducted in the Venice lagoon. Leach *et al.* (2013) recorded significantly more cognitive then normative learning in their study on collaborative partnerships, and normative learning was absent from the adaptive co-management learning study conducted recently by Baird *et al.* (2014). What reasons can be given for its scarcity? One possibility is proffered here, based on the evidence that an actor involved as a concerned individual in a personal capacity was significantly more likely to change their priorities than any other actor, particularly an expert or business actor. This implies that some actor types may be more willing or open to questioning or changing their priorities, with those who consider themselves knowledgeable in the field being more reluctant to admit to a value shift (Haug *et al.* 2011). However, this does not explain why business actors experienced the least normative learning. An alternative explanation is that norms and belief systems underlie our positions and actions and the more "face" to lose the more defensive we are of our beliefs, especially in a professional situation dealing with actual policy concerns. This assumption contrasts the one made by Baird *et al.* (2014) that time is the major factor in producing normative learning (a finding not replicated here), and brings up questions regarding how unfeasible it is to expect normative learning in situations where, despite a deliberative setting, there is no practical solution to getting around defences that protect actor interests. Moreover, normative learning does not imply an improvement, just a change, and the adjustment of norms and values could also be the result of persuasion tactics by more powerful actors (Haug *et al.* 2011). In light of all this, perhaps an improvement in understanding of the facts, in the understanding others' mind-sets, and a growth of trust is enough to address the complexity and uncertainty that shroud climate policy problems and should bare more of the focus in policy research.

The second point of discussion reflects on the use of the conceptual framework to model reality and predict learning effects. The positioning of experiments at the science-policy nexus to study learning was arguably novel (compared to recent learning studies that follow the general trend of analysing participatory settings) but it stemmed from an assumption that experiments most likely embody the technocratic model. Therefore it was a surprise that not more cases fitted this type, especially when one of the criteria for the cases was that they test an ecosystem response. It was the least common design in most of the indicators measured, with far more experiments having boundary type characteristics than assumed. This could reflect the nature of the Dutch 'polder model', which embodies consensual negotiated knowledge (Owens *et al.* 2004) and that experiments play more of a role in producing socially robust knowledge than previously thought (Nowotny 2003). With post-normal science considered a more superior mode of knowledge governance for wicked problems than traditional forms (Petersen *et al.* 2011) this is a strong win for proponents of using the experimental method to test policy concepts.

It cannot be ignored, however, that the advocacy experiment was also common. This implies that policy actors may be using the tool as a way to skim over ethical and political choices by framing issues as technical and requiring experimental testing without questioning norms or power structures (Owens *et al.* 2004). However, despite the indicators for power structure and actor inclusiveness being typically advocacy type, there were no deep underlying divisions detected in the cases. The information transmission in all cases was generally considered open and transparent by participants, and no claims were made about crucial parties being omitted from the process, nor that the experiment was conducted to defer a policy decision (5 participants claimed this as the reason for the experiment, and 4 were policy actors). Therefore, although a fair portion of the cases were advocacy types, they were not of a sinister kind. Advocacy types were also significantly less likely to be controversial, have legal barriers, or cause concern in the surrounding area compared to the other types, indicating that policy actors may use advocacy experiments for structured, uncontentious problem issues (as opposed to technocratic types, as expected by Owens *et al.* 2004).

Finally, a few points can be made about how the design influenced learning outcomes. The results confirmed that technocratic experiments produce significantly more cognitive learning, and it is somewhat surprising that boundary types produced less than advocacy types. Is it possible that transparent, deliberative, and inclusive policy processes supress knowledge acquisition? Three explanations are possible. First, boundary experiments have a range of actor types and the lack of prior knowledge or the analytic sophistication of an expert could drag down the cognitive learning scores. However, Leach *et al.* (2013) counters this proposition, as in their study those with lower competence actually learned more. A second explanation could follow van Eeten's dialogue of the deaf metaphor, where long standing policy controversies between sides in the policy community render them unable to resolve issues using the facts (Eeten 1999). People talk but do not listen, so they cannot learn. However, checking the extent to which the experiments were dealing with controversial issues, results show that although boundary experiments were more controversial than advocacy experiments, the technocratic experiments were significantly more controversial. Therefore, this explanation is possible but weak. The third explanation involves the offsetting of instrumental vs reflexive knowledge. Bivariate statistics show that experiments with a high number of participants that discussed goals of the project had significantly less cognitive learning. Somehow, the more an experiment engages participants in the discussion on goals, the less knowledge and understanding produced. This finding has implications for the somewhat clear cut argument in policy sciences that combining the consideration of both fact and values in an experiment would be of benefit (Dryzek 1987; Fischer 1995), and further research (preferably of an in-depth, qualitative nature) needs to be conducted to shine light on this trade-off.

8 Conclusions

The intention of the article is to present a conceptual framework siting policy experiments at the science-policy interface and linking their design features to preferred learning outcomes, in order to answer the research question: t*o what extent can policy learning be explained by an experiment's institutional design?* The results show that design goes far in explaining what sort of and how much learning is produced.

For the purposes of this research, policy experiments are defined as *a temporary, controlled field-trial of a policy-relevant innovation that produces evidence for subsequent policy decisions.* Eighteen cases were identified out of an initial 174 innovative initiatives in Dutch climate adaptation, indicating that policy experimentation is actually rather rare in climate governance. The cases tested policy concepts in coastal, freshwater, and flooding, drought management. Out of 18 cases, four fit the description of a technocratic ideal type, which are those that separate science from power. Seven experiments fit the boundary type, which embody transparent, deliberative, and inclusive processes; and seven fit the advocacy ideal type, which are driven by policy actors and restrict information exchange and authority distribution. The technocratic experiments produced significantly high cognitive learning, some normative learning, and some relational learning, partially meeting the hypothesis. Boundary experiments produced the lowest cognitive learning, and highest amounts of normative and relational learning. Their levels of relational learning were significantly higher than others but the low cognitive learning was a surprise. Finally, meeting expectations, the advocacy experiments produced low levels of normative and relational learning, but the cognitive learning was higher than expected. Cognitive and relational learning scores were significant among the groups, confirming that for these effects design has an effect on how much learning is produced.

The discussion shows that experiments in Dutch climate adaptation are more likely to entwine the fields of policy and science and produce policy-relevant knowledge than be set apart from the policy process and produce solely scientific evidence. An experiment turns out of be a successful method of involving non-state actors in the policy process, although their authority to make decisions is considerably limited. Based on the learning results, it is concluded that initiators need to think carefully when designing an experiment, in particular who is involved, how much power they have, and what information is transmitted and who to, because the findings show that when designing experiments there is essentially a trade-off between building trust among participants and increasing their knowledge acquisition. How to get participants to change their values and beliefs is another question, as the results show that experiments do not tend to enable this learning effect.

A study on learning would not be complete without the requisite limitations section. Obviously there are caveats to be drawn regarding how valid learning data is when gathered *ex post*, via self-reported learning methods. This is why an attempt to be as standardised and thorough as possible was made, with closed questions and a range of question types; for example, asking factual questions about a participant's authority, as well as about their opinion on how open the experiment was to outsiders. The learning questions were based on previously published studies, although had to be made intentionally vague so to capture the content of different experiments. Quantitative analysis paints with broad strokes and picks up patterns across many cases, but qualitative research is superior to find underlying dynamics and reasons for learning. Finally, despite a comprehensive search for experiments, it would be unwise to extrapolate findings to experiments in policy at a whole, due to the purposive, snow-ball sampling strategy.

Further research into the relationship between experiments and learning could examine how the policy network reacts to these sorts of initiatives. Do the ideal types create different perceptions of how credible, salient, or legitimate the experiment's evidence is? If an experiment has high learning scores does this translate into impact on the wider policy network? Further work could also push forward the connections identified in this article or reapply the framework to individual cases. The relationship between learning and experimentation is often assumed but exploration is in its infancy, if it grow up then climate governance will be the better for it.

References

Anderies, J.M. & Janssen, M., 2013. Robustness of Social-Ecological Systems: Implications for Public Policy. *Policy Studies Journal*, 41(3), pp.513–536.

Armitage, D., Marschke, M. & Plummer, R., 2008. Adaptive co-management and the paradox of learning. *Global Environmental Change*, 18(1), pp.86–98.

Baird, J., Plummer, R., Haug, C., and D. Huitema. 2014. Learning effects of interactive decision-making processes for climate change adaptation. *Global Environmental Change*, 27, pp.51–63.

Berkhout, F. *et al*., 2010. Sustainability experiments in Asia: innovations shaping alternative development pathways? *Environmental Science & Policy*, 13(4), pp.261–271.

Campbell, D.T. (1998). The Experimenting Society. In Dunn, W. (ed.), *The Experimenting Society: Essays in Honor of Donald T. Campbell. Policy Studies Review Annual volume 11* (p. 35). New Brunswick, New Jersey: Transaction Publishers.

Caspary, W. (2002). *Dewey on Democracy.* Cornell University Press, USA.

Castán Broto, V. & Bulkeley, H., (2013). A survey of urban climate change experiments in 100 cities. *Global environmental change : human and policy dimensions*, 23(1), pp.92–102.

Crona, B.I. & Parker, J.N., (2012). Learning in Support of Governance : Theories , Methods , and a Framework to Assess How Bridging Organizations Contribute to. , 17(1).

Cundill, G. & Rodela, R., (2012). A review of assertions about the processes and outcomes of social learning in natural resource management. *Journal of environmental management*, 113, pp.7–14.

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

Duijn, M. 2009. Embedded Reflection on Public Policy Innovation. Uitgeverij, Delft, Netherlands.

Eeten, M.J.G. Van, 1999. "Dialogues of the deaf" on science in policy controversies. *Science and Public Policy*, 26(3), pp.185–192.

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

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

Gerlak, A. & Heikkila, T. (2011). Building a theory of learning in collaboratives: Evidence from the Everglades restoration program. *Journal of Public Administration Research and Theory, 21*, 619–644.

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

Grin, J. & Loeber, A. (2007). Theories of policy learning: Agency, structure, and change, In Fischer, F., Miller, G.J. & Sidney, M.S. (eds.), *Handbook of Public Policy Analysis: theory, politics, and methods* (pp.201-219). Florida, USA: Taylor & Francis Group.

Haug, C., Huitema, D. & Wenzler, I. (2011). Learning through games? Evaluating the learning effect of a policy exercise on European climate policy. *Technological Forecasting and Social Change, 78*(6), 968–981.

Huitema, D., Mostert, E., Egas, W., Moellenkamp, S., Pahl-Wostl, C., Yalcin, R., 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.

Huitema, D., Cornelisse, C. & Ottow, B. (2010). Is the jury still out? Toward greater insight in policy learning in participatory decision processes—the case of Dutch citizens' juries on water management in the Rhine Basin. *Ecology and Society, 15*(1), 16.

Kuks, S. (2002). The Evolution of the National Water Regime in the Netherlands. University of Twente (UT).

Leach, W.D., Weible, C.M, Vince, S.R., Siddiki, S.N. & Calanni, J.C. (2014). Fostering learning through collaboration: knowledge acquisition and belief change in marine aquaculture partnerships. *Journal of Public Administration Research and Theory, 24*(3), 591-622. doi:10.1093/jopart/mut011.

Lee, K.N., 1999. Conservation Ecology: Appraising Adaptive Management. *Ecology and Society*, 3(2).

McFadgen, B. and Huitema, D., in review. Experimentation and learning. The design of policy experiments and their learning effects, a conceptual framework and application to a case study from the Netherlands. *Ecology and Society.*

Meijerink, S. & Huitema, D., 2007. Understanding and managing water policy transitions: a policy science perspective. pp.23–36.

Mostert, E., Pahl-Wostl, C., Rees, Y., Searle, B., Tàbara, D. & Tippett, J. (2007). Social learning in European river-basin management: barriers and fostering mechanisms from 10 river basins. *Ecology and Society, 12*(1), 19.

Munaretto, S. & Huitema, D., 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).

Muro, M. & Jeffrey, P. (2012). Time to talk? How the structure of dialog processes shapes stakeholder learning in participatory water resources management. *Ecology and Society, 17*(1), 3.

Nowotny, H. 2003. Dilemma of expertise: Democratising expertise and socially robust knowledge. 30(3), pp.151–156.

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

Owens, S., Rayner, T. & Bina, O., 2004. New agendas for appraisal: reflections on theory, practice, and research. *Environment and Planning A*, 36(11), pp.1943–1959.

Petersen, A.C. *et al.*, 2011. Post-Normal Science in Practice at the Netherlands Environmental Assessment Agency. *Science, technology & human values,* 36(3), pp.362–388.

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

Poteete, Amy, Marco Janssen, and Elinor Ostrom. 2010. *Working Together: Collective Action, the Commons, and Multiple Methods in Practice.* Princeton, NJ: Princeton University Press.

Reed, M., Evely, A.C., Cundill, G., Fazey, I.R.A., Glass, J., Laing, A., Newig, J., Parrish, B., Prell, C., Raymond, C., Stringer, L., 2010. What is social learning? *Ecology and Society.* 15.

Rodela, R. (2011). Social learning and natural resource management: the emergence of three research perspectives. *Ecology and Society, 16*(4) 30.

Sabatier P. 1988. An advocacy coalition framework of policy change and the role of policy-oriented learning therein, *Policy Science*. 21 (2–3) 129–168.

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

Steffen, W. (2009). Interdisciplinary research for managing ecosystem services. *Proceedings for the National Academy of Sciences, 106*(5), 1301–1302.

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.* New Brunswick, USA: Transaction Publishers.

Voß, J. & Bornemann, B. (2011). The politics of reflexive governance: challenges for designing adaptive management and transition management. *Ecology and Society, 16*(2), 9.

Annex A

	Indicator	Explanation	Settings
Position rules			

1A	*Respondents in stakeholder role*	Extent participants identify themselves as stakeholders.	BIT= >50%
			AIT= 20-50%
			TIT= <20%

1B	*Initiator type*	What actor type initiated the experiment.	BIT= collaboration between two or more actor types.
			AIT= policy actor
			TIT= expert actor

1C	*Facilitator involvement*	How many participants recall the presence of an independent facilitator during the experiment.	BIT= >66% recall
			AIT= 20-65% recall
			TIT= <20% recall.

Choice rules			

2A	*Design decisions*	Analysis of how the authority to make decisions on design was distributed among the actor types.	BIT= collaboration between more than two parties.
			AIT= policy actor only.
			TIT= expert actor only; expert driven

2B	*Monitoring decisions*	Analysis of how the authority to make decisions on monitoring and implementation was distributed among the actor types.	BIT= collaboration between more than two parties/ pol with other parties. AIT= policy actor majority TIT= expert majority; collaboration between expert and policy.
2C	*Evaluation decisions*	Analysis of how the authority to make decisions on evaluation was distributed among the actor types.	BIT= collaboration between more than two parties/ pol with other parties. AIT= policy actor majority TIT= expert majority; collaboration between expert and policy.
Information rules			

3A	*Contribute to goal discussion*	Percentage of participants that contributed to the discussion on project goals.	BIT= > 75%
			AIT= 30-75%
			TIT= <30%

3B	*Lay knowledge contribution*	Percentage of participants that (solely) contributed lay knowledge.	BIT= >50% (at least one solely)
			AIT= 30-50%
			TIT= <30%

3C	*Scientific knowledge contribution*	Percentage of participants that contributed scientific knowledge.	TIT= >50%
			BIT= 25-50%
			AIT= <25%

3D	*Sufficient information sharing*	To what extent participants were satisfied with how much information was shared within the group.	BIT= >75% agree (everyone found sufficient).
			AIT= >50% neutral or disagree (minority found sufficient).
			TIT= remaining (majority found sufficient).

3E	*Information openly shared*	To what extent participants found the process of sharing information open among the group.	BIT= >75% agree (everyone found sufficient).
			AIT= >50% neutral or disagree (minority found sufficient).
			TIT= remaining (majority found sufficient).

3F	*Sufficient personal contact*	To what extent participants felt there was sufficient personal contact among the group.	BIT= >75% agree (everyone found sufficient).
			AIT= >50% neutral or disagree (minority found sufficient).
			TIT= remaining (majority found sufficient).

3G	*Outsiders informed of progress*	How regularly non-participants were informed of the experiment's progress.	BIT= >75% state often informed.
			AIT= >50% state irregularly/not informed.
			TIT = remaining state somewhat informed.

Aggregation rules	*How decisions are made- unilaterally or shared and by who if the former-*		
4A	*Amend power*	How decisions are made about amending the experiment. Whether the power is shared among the group or decisions taken by initiators. Classification considers dominant actor type and who initiated.	BIT= broad actor type sharing decision power. AIT= dominant policy actors sharing power, or policy actors as initiators. TIT= dominant expert actors sharing power, or expert actors as initiators.
4B	*End power*	How decisions are made about ending the experiment. Whether the power is shared among the group or decisions taken by initiators. Classification considers dominant actor type and who initiated.	BIT= broad actor type sharing decision power. AIT= dominant policy actors sharing power, or policy actors as initiators. TIT= dominant expert actors sharing power, or expert actors as initiators.
Pay off rules			
5A	*How costs were distributed*	Look into how the costs of the experiment were paid; to what extent participants were	BIT= costs were shared. AIT= a participant's organization paid all

		expected to “buy-in” to the experiment.	costs or no costs. TIT= no clear distinction; some paid, some shared.
Boundary rules			

6A	*Inclusiveness*	Inclusiveness considers the breadth of actors, broadest being all five actor types. If there is no dominant actor, then four actor types is also broad. Dominant actor is one that equals or is more than other types together.	BIT= All five actor types; or four with no dominance.
			AIT= Dominant policy actor, 2-4 other types.
			TIT= Dominant expert

6B	*New actor network*	To what extent the participants have met each other before.	BIT= >66% never met before or met less than half. AIT= >66% know everyone or met more than half. TIT= remainder.
6C	*Describe other participants*	Participants asked to describe other actors in the group.	BIT= enthusiastic for a new idea. AIT= actors needed to change policy/ a good link between science and policy. TIT= group of experts.
6D	*Openness to new participants*	To what extent new participants can join the project once it has started.	BIT= all new actors welcome or mostly welcome. AIT= no one welcome or marginally welcome. TIT= mixed.

6E	*Gaining access*	How participants entered the experiment. Participants can either be invited in, involved because they are part of the organizing team, or because they requested involvement from the organisers. The most common method is used for classification, except if anyone requested involvement, then a BIT classification is used.	BIT= Requested involvement
			AIT= Organiser/obliged
			TIT= Invited in